Title
Investigations of Color Categorizations and Their Evolution

Author
Joe, Kirbi Caryn

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Investigations of Color Categorizations and Their Evolution

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical Behavioral Science

by

Kirbi Caryn Joe

Dissertation Committee:
Professor Louis Narens, Chair
Professor Emeritus A. Kimball Romney
Professor Zygmunt Pizlo

2020

DEDICATION

To Mom and Dad

My biggest supporters and constant examples of Christ-likeness

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

They say that it takes a village, which I know now to be one of the truest statements ever spoken. I would not be here without my village, to whom I want to extend my deepest thanks and gratitude.

First and foremost, I want to thank Kimberly Jameson, to whom I am indebted to for this entire experience. She has been my biggest cheerleader, mentor, and guide ever since she took me under her wing as a wayward undergraduate and all the way through my doctoral program. I am so grateful to have been under her tutelage as an undergraduate research assistant, graduate student, and colleague. And to Louis Narens, thank you for constantly moving the goal posts back so that I had no choice but to finish this degree. I would not have been able to make it through grad school without their unwavering guidance and support.

Thanks to my parents, Tommy and Doreen, and my brother, Kory, for being the most supportive and loving family I could have asked for. They have been my prayer warriors through the ups and downs of these last few years and have never ceased to point me back to Christ. I am so thankful for the example that they've shown me of how to work heartily unto the Lord and how to continue trusting in His sovereign plan even when it's hard. Love you guys lots.

To my Berean family, thank you for being my spiritual home and community for the last eight years. Thank you to the church leaders for teaching and guiding me through the Word of God and for being strong examples of faith. Thanks to all of my friends who have listened to me share through my struggles with school and have prayed for and encouraged me to remain anchored in Christ.

My time in the IMBS has been so special and I am immensely grateful for all of the wonderful students, faculty, and staff that I have had the pleasure to be instructed by and collaborate with. Specifically, I want to thank my committee, Louis Narens, Kim Romney, and Zyg Pizlo for their invaluable feedback and suggestions on my research. Also, I want to thank Nikhil Addleman, Lucila Arroyo, Calvin Cochran, and Maryam Gooyabadi for being the best cohort ever and making grad school that much more enjoyable.

Financial support from the Institute for Mathematical Behavioral Sciences, the School of Social Sciences, and the International Foundation for Research in Experimental Economics is gratefully acknowledged. Part of this dissertation is a reprint of the material as it appears in the *Journal of the Optical Society of America A*. A co-author listed in this publication directed and supervised research which forms the basis for the dissertation. Additional thanks to Maryam Gooyabadi who co-authored some of this work with me and who has been a true collaborator for many years.

Lastly and most importantly, I am most thankful for my Lord and Savior, Jesus Christ, to whom I owe my life. All glory be to Christ.

CURRICULUM VITAE

Kirbi Caryn Joe

EDUCATION

Doctor of Philosophy in Mathematical Behavioral Science **2020**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Mathematics **2016**
University of California, Irvine *Irvine, CA*

RESEARCH EXPERIENCE

Graduate Researcher **2016–2020**
University of California, Irvine *Irvine, California*

Undergraduate Research Assistant **2015–2016**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2016–2020**
University of California, Irvine *Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Joe, K., Gooyabadi, M. (2020). A Bayesian nonparametric mixture model for studying universal patterns in color naming. *Submitted.*

Jameson, K. A., Satalich, T. A., **Joe, K. C.**, Bochko, V. A., Atilano, S. R., & Kenney, M. C. (2020). Human Color Vision and Tetrachromacy. *Cambridge University Press.*

Gooyabadi, M., **Joe, K.**, & Narens, L. (2019). Further evolution of natural categorization systems: an approach to evolving color concepts. *JOSA A*, 36(2), 159–172.

SELECTED CONFERENCE PRESENTATIONS

Further Evolution of Natural Categorization Systems **June 2019**
25th Anniversary GWCSS Alumni Conference at the Santa Fe Institute

Investigating Variations in Color Perception among Artist and Non-artist Participants with Varying Photopigment Opsin Genotypes (Poster) **March 2019**
Bench-to-Bedside Conference at the Gavin Herbert Eye Institute

Quantifying Diachronic Change in Category Meaning Using Evolutionary Models of Learning **November 2018**
Formal Modeling and Analysis of Color Categorization: Innovations and Insights Since Berlin and Kay (1969)

WORKSHOPS

Graduate Workshop in Computational Social Science II **June 2019**
Santa Fe Institute

Graduate Workshop in Computational Social Science **June 2018**
Santa Fe Institute

Internet-based Data Collection and Analysis in Judgment and Decision Making **September 2017**
Universität Konstanz

GRANTS, HONORS, AND AWARDS

2nd Place Poster Award ($300), Bench-to-Bedside Conference, GHEI
Small Grant ($3000), International Foundation for Experimental Economics (IFREE)
Multidisciplinary Design Program ($2400), UROP, UC Irvine
Christian Werner Fellowship ($7000), School of Social Sciences, UC Irvine
Associate Dean's Fellowship (in place of TAship), School of Social Sciences, UC Irvine
Summer Research Support ($2500), IMBS, UC Irvine
Phi Beta Kappa, University of California, Irvine
Campuswide Honors Program, University of California, Irvine
Dean's Honor List, School of Physical Science, UC Irvine

SOFTWARE

ColorSims 2.0 `https://github.com/kirbijoe/colorsims2`
Python-based software for the evolving of color naming systems using empirical seed data taken from the World Color Survey.

ColorBBDP `https://github.com/kirbijoe/colorbbdp`
Python-based software for using the Beta-Bernoulli Dirichlet Process Mixture Model to perform a clustering of the World Color Survey color naming data.

ABSTRACT OF THE DISSERTATION

Investigations of Color Categorizations and Their Evolution

By

Kirbi Caryn Joe

Doctor of Philosophy in Mathematical Behavioral Science

University of California, Irvine, 2020

Professor Louis Narens, Chair

This dissertation applies various mathematical and computational methods to investigate properties of color naming systems and their evolution. An advantage of using quantitative analyses to study behavioral topics is the ability to concretely define complex processes and phenomena in a precise and controlled manner. In this manuscript, I present three studies which utilize computational models and mathematical analyses to examine features of both static and evolving color naming systems. The first two studies use evolutionary game theory to simulate the evolution of categorization systems modeled using data from the World Color Survey. The first study investigates the evolution of these naturally occurring systems under a simple communication mechanism and the second study explores how the introduction of non-salient regions in the color space affects the evolutionary process. The third study applies a Bayesian nonparametric mixture model to the World Color Survey data in order to reveal universal patterns in color naming across different languages. This dissertation argues for the effectiveness of applying quantitative methods to behavioral data sets, such as the World Color Survey, and explores the types of inferences which can be drawn from these results. Finally, preliminary results from a pilot study examining the correlation between minimized flicker settings and observer phenotype are also presented.

Chapter 1

Introduction

Color is a seemingly simple and intuitive concept that is instilled in us from an early age. The ability to identify an object's color is a skill that children are expected to master within the first couple years of life. Though the act of identifying the color "blue", "red", or "green" might seem like a cognitively simple task, the processes which must occur in order for rays of light to be translated into lexical terms which are communicable across observers are immensely intricate and complex.

The physical quantity of color can be captured using the range of wavelengths within the electromagnetic spectrum which are visible to the human eye, 340–740 nm. Shorter wavelength light produces colors English speakers generally refer to as "blue" or "purple" and longer wavelengths produce "reddish" colors. When a human experiences the sensation of color, it is caused by light emitted from an illuminant, such as the sun or a bulb, reflecting off of the surface of objects and entering the observer's eye. Based on the wavelength of the light which enters the eye, some proportion of the cones in the eye's retina will become stimulated. Stimulated cones then send signals to the brain which undergo several more layers of processing before the observer perceives the color of the object. Once the color is

perceived, there are additional levels of processing that must occur in order for that color to be associated with a name. What's more, the lexical term that an observer assigns to a color must have shared meaning among the population in order for it to be successfully communicated. Therefore, social learning and coordination are necessary for the formation of these color-based conventions.

Due to its varied and complex nature, the study of color spans a wide variety of disciplines. Color processing at both the level of the eye and brain have been studied extensively by psychologists and vision scientists [49, 90, 91, 57, 115]. The physical properties of color have been examined by physicists [88, 94]. Linguists and anthropologists have theorized how populations of observers come to a shared understanding of color in the context of their language and culture [10, 109, 78, 73]. The study of color has even been extended to examine its influence over our emotions and the actions that we take [3, 134, 22, 42, 2, 97]. While it is clear that there are a plethora of questions that could be asked with regards to humans and their relation to color, this dissertation specifically endeavors to discuss the features and evolution of color naming systems. This discussion is impossible without acknowledging the seminal work of Brent Berlin and Paul Kay. Though the study of color naming long predates Berlin and Kay, much of the recent work regarding color naming and its evolution has been built on the foundation of their work, starting with their 1969 book, *Basic Color Terms: Their Universality and Evolution* [10].

1.1 Basic Color Terms

When Berlin and Kay published *Basic Color Terms* [10], they revived the color naming debate and sparked a renewed interest in the study of color naming in academia. They drew two primary conclusions about the acquisition and evolution of color terms: (i) that there was a limited set of color words from which all languages derived their color terms

and (ii) that languages acquired these terms in a relatively fixed order. These terms were coined *basic color terms* (BCTs)—i.e. the smallest set of color category names with which a person could name the entire color space. BCTs were determined according to the following semantic and syntactic criteria:

1. It is monolexemic (blue, not blue-green).
2. It is monomorphemic (blue, not bluish).
3. It is not contained in another, broader color category (scarlet is a type of red).
4. Its use is not limited to a narrow class of objects (blond is usually restricted to hair and beer colors).
5. It has to be psychologically salient to the general population ("the color of my car" is not salient to most people).

There are 11 BCTs identified for English: white, black, red, yellow, green, blue, orange, brown, purple, pink, and gray. These 11 terms were also theorized to be the limited set from which all languages draw their color terms. Applying these criteria to data collected through "published sources and personal communication with linguists and ethnographers [with] specialized knowledge of the languages in question" [10], Berlin and Kay identified the BCTs for 78 languages and for 20 additional languages which they collected in their own empirical study. Through this analysis, they concluded most languages have between 2 and 11 BCTs.

In addition to defining this constrained set of possible BCTs, Berlin and Kay found that languages followed a fixed evolutionary pattern. They organized this pattern into a seven stage evolutionary path, positing that languages acquired new color terms in the following order:

1. Stage I: Dark/cool (black) and light/warm (white)
2. Stage II: Red
3. Stage III: Either green or yellow
4. Stage IV: Both green and yellow
5. Stage V: Blue
6. Stage VI: Brown
7. Stage VII: Brown, purple, pink, and gray

In response to new, emerging research, this evolutionary hierarchy was altered in 1999 into the five stage hierarchy depicted in Figure 1.1 [69]. This new hierarchy still proposed that languages acquired color terms in a fixed order but allowed for several possible paths by which languages could incorporate these new terms into their existing system.

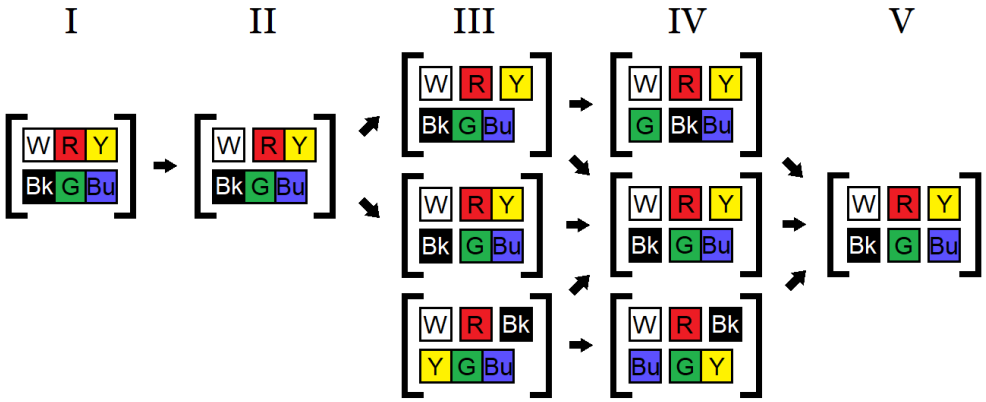

Figure 1.1: Evolutionary stages and their corresponding basic color term (BCT) systems using English term equivalents. Stages are determined by the number of BCTs in the system. Colored blocks that are connected represent a single term which encompasses all of those color regions. For example, Stage I is a 2 BCT stage where one term refers to the white/red/yellow region of the color space and the other term refers to the black/green/blue region.

Berlin and Kay's work is noteworthy because it presented arguments in favor of the view that color naming is driven by universal principles, not relative, culture-specific phenomena.

By lending such explicit support for the *universalist* view, they reignited the color naming debate and prompted decades worth of new research supporting, challenging, and amending their original hypotheses.

1.2 World Color Survey

As part of an effort to provide evidence of their claims regarding BCTs, Berlin and Kay conducted an empirical study to collect color naming data from a small sample of speakers [10]. This study was met with several criticisms. Some of the critics argued that (i) the sample was too small (three or fewer speakers per language), (ii) the speakers were bilingual in their native tongue and English, (iii) the data was collected in Northern California instead of their native communities, and (iv) the languages represented by the sample were mostly from industrialized societies [65]. In response to these critics, Berlin and Kay embarked on a project called the *World Color Survey*. They addressed the criticisms of their previous empirical work by collecting a much larger sample which consisted of monolingual speakers from pre-industrial, tribal languages worldwide. The final data set was comprised of 110 languages, each with 24 participants on average (modal number was 25) [23].

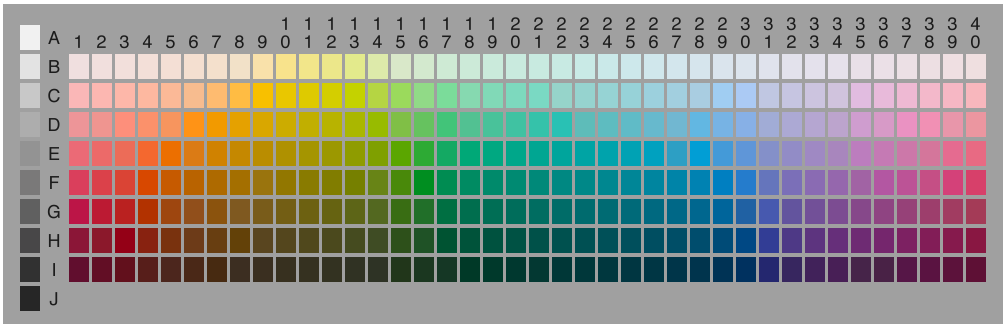

Figure 1.2: The set of 330 Munsell color chips used in the World Color Survey.

Participants of the World Color Survey completed two color naming tasks: the *mapping task* and the *naming task*. The *mapping task* was a focal identification procedure. The facilitator would present a word to the participant (from a list of BCTs that was previously

elicited from the participant), then the participant would indicate which color chip in the Munsell color grid (Figure 1.2) was the best exemplar of that word. The *naming task* instead presented participants with a color chip and prompted them for a name for that color. For this task, participants were asked to name all 330 color chips in Figure 1.2 one at a time in a fixed, random order. Once compiled, the *naming task* data for a participant creates a full partition of the color space. Kay and colleagues concluded that the findings from the World Color Survey were in agreement with the original claims made by Berlin and Kay [66]. Particularly, they observed variation in the number of color terms used to partition the color space across the collected languages. They also found that color terms were systematically present or absent from a population's color lexicon depending on the number of basic color terms in the lexicon. These findings were determined to be consistent with Berlin and Kay's hypothesis about the existence of a shared, universal evolutionary path. That is, it appeared that languages did acquire color terms in a fixed order and the observed variation could be attributed different languages being at different stages of evolution.

Since its release, the data from the World Color Survey has been used extensively in various studies to study the factors that might have manifested the observed cross cultural differences in color naming [108, 111, 77, 78, 15, 4, 31, 30, 32, 39]. In the same vein, the World Color Survey data set is used to both inform the mathematical models presented in this dissertation and to assess how well the theoretical results obtained from these models compare to the empirical reality.

Chapter 2

Further Evolution of Natural Categorization Systems

2.1 Introduction

A major area of interdisciplinary study investigates how words and signals acquire meaning. This is a central and heavily investigated issue in the humanities (philosophy and linguistics), the social sciences (psychology and anthropology), and engineering and computer science (robotics and artificial intelligence). This chapter investigates a special case: the evolution of color concepts in languages of non-industrialized, isolated linguistic communities.

Color naming has a long history in academia, beginning with seminal work on ancient Greek color terminology by Gladstone in 1858 [36] that was extended by other 19th century researchers to additional ancient languages. The subject gained increasing recognition in 1969 due to the seminal study of Berlin and Kay's, *Basic Color Terms: Their Universality and Evolution* [10] (see Section 1.1), and later from the *World Color Survey* (WCS) by Kay, Berlin, Maffi, and Merrifeld [66] (see Section 1.2). This study provides a new methodology

for analyzing data from the WCS based on evolving naming strategies through evolutionary game theory.

2.1.1 Rationale for Game Theoretic Approach

There is a precedence of using game theory to understand the evolution of meaning in linguistic and signaling systems. There are many approaches to this subject [125, 126, 122, 46, 25, 81, 21, 5], and several have been used for color naming. Other approaches have employed simulation-based methods specifically on the WCS data to look at features of color naming systems (e.g. optimal partitions of the Munsell color space [111] and perceptual constraints leading to the emergence of universal patterns [4]).

We follow this path and consider color names as conventions [121], and use the *discrimination-similarity game* and the *2-player teacher game*, developed in Komarova, Jameson, and Narens [73] as a way to evolve population color naming strategies. The dynamics of these games are based on a form of reinforcement learning. Unlike some other evolutionary color naming models which emphasize supposed properties of human color vision (e.g. the Hering primaries of red, blue, green, and yellow lights having special perceptual properties and salience), the game dynamics are based only on primitive ideas about communication and the agents' ability to discriminate one colored stimulus from another.

The WCS revealed a highly restricted pattern of naming strategies across linguistic communities [40]. A number of theories have been developed to explain this pattern. The idea of those who prescribe to the universalist view, is that each community is in a particular stage of color naming evolution. The overall evolutionary process of color naming in a certain population can be understood by analyzing the increasing complexities of the naming strategies in terms of how populations partition the color space into a set of concepts called "basic color categories" (BCCs) using "basic color terms" (BCT)—the minimum number of color

terms needed for each linguistic population in order for a speaker to name every possible color (see Section 1.1).

Our analysis and goal for the WCS data is different. Like Berlin and Kay, we take each linguistic community to be at a particular stage of color category evolution, but we do not then use another unrelated language with an additional BCC/BCT to predict the next, more complex stage of that linguistic community evolution. Instead, we ask: how well evolved, *as a communication system*, is the community's naming strategy, and what would its future evolution look like if its population communicated freely about color?[1] This is modeled in the evolutionary dynamics by using individual data from the WCS to model agents in a population and having them play the communication game repeatedly until an equilibrium strategy (i.e. a color naming convention) arises and remains stable.

This evolutionary approach is particularly useful for two reasons. First, much of the data available on color naming, including the WCS, is cross-sectional. To investigate the evolution of naming systems, diachronic data would be most ideal because it would provide real observations of how a community's naming system changes over time. However, data of this type is extremely sparse, and now impossible to attain for linguistic communities that have gone extinct or are composed of bilingual speakers. Therefore, simulations can generate psuedo-diachronic data through repeated agent interaction modeled after human communication. Other approaches that attempt to approximate features of naming systems involve using basics of modern semantics to retrieve past color systems [12, 13, 14]. This approach attempts to approximate past systems using current categorization schema while our attempts aim at understanding features of stable, convergent systems. Evolving the cross-sectional data in this way provides a unique perspective on possible ways which color

[1]Of the many possible factors that could play a role in the evolution of color categories, the purpose of this method is to solely evaluate agent communication mechanisms. Therefore, the model that we present is not a predictive model and its interpretation does not serve as a foretelling of what these color naming systems might look like in the future. This method is simply a proof of concept that allows us to investigate color category evolution in a controlled manner so that we may identify key features and concepts of category evolution that could then inform future empirical studies of real populations.

naming systems evolve, which has not been done previously.

Second, if the evolutionary dynamics prove successful, it opens up possible avenues for further exploration. For example, the pioneering work of Berlin and Kay [10] grouped languages into evolutionary stages based on the number of BCCs/BCTs they contained. Later, they applied their linguistic approach to determine the number of BCTs in the 110 languages collected in the WCS [66]. Fider, Jameson, and Komarova [31] confirmed many of the conclusions made with regards to the WCS, but instead determined the number of BCTs using a computational method which requires no knowledge of semantics of a language's color words. However, both these methods did not provide information on where languages lie within a stage—that is, did a language recently enter a stage by acquiring a new BCT? Or is a non-basic term growing in prominence within a language, positioning it on the cusp of entering the next stage? By evolving each linguistic group's data independently to a stable equilibrium, we can potentially (i) observe the evolution of individual color concepts and draw conclusions about terms that are "falling in" or "out" of basicness, and (ii) propose a new grouping criterion for comparison across naming systems that provides a richer understanding of a language's evolutionary path.

2.1.2 Theories of Color Categorization

Berlin and Kay's theory for color categorization and linguistic evolution assumes that every color always possesses a unique color name. This is reflected in the data they collected for their 1969 book [10] and in the WCS [66], where participants assigned names to all stimuli in the color space. With a fully named stimulus space, Berlin and Kay were able to partition the color space into color categories and identify color terms that were "basic" for each population. This is known as the *partition principle* [69]. However, this procedure could yield regions in the color space for which there are multiple BCTs—participants from the

same population using different words to describe the same color region. In other words, since the task was forced naming, populations may not have an agreed upon BCT for some color chips. We consider these regions with low intra-population consensus to be gaps that could hinder effective communication about color. An alternative view to the *partition principle* asserts that the whole color space does not need to be named (e.g. see [76]). Rather, the assignment of BCTs is limited to focused, salient regions of the color space, which gives rise to empty or ambiguously named areas in the color space.

According to the *partition principle*, the introduction of a new term can only arise by splitting an existing category. Japanese provides a good example of this with the single term *aoi* for blue and green hues that split in the 14th and 15th century A.D. into *aoi* for blue and *midori* for green [123, 124]. In more modern times, further splitting has been noted in a recent study [74] demonstrating that Japanese has two BCTs for blue: *kon* for "indigo"/dark blue and *mizu* for "water"/light blue. This change has been observed since the study of Japan's number of BCTs in the 1980s [133]. Similarly, Russian's *blue* category—originally called *sinij*—split into two categories, *sinij* for darker blues and *goluboj* for lighter blues [99, 100]. This theory of color category evolution is called the *Partition Hypothesis*. In addition to this version of the *Partition Hypothesis* which implies *category fission*, another version claims that the color space can divide by a new term inserting itself. For example, modern English speakers augment the modern BCT inventory by commonly using "*turquoise*" to describe the *blue-green* boundary [92]. This augmented theory is called *cross-boundary insertion* or the *category insertion* version of the *Partition Hypothesis*.

The alternative to the *Partition Hypothesis*, which in this paper will be simply referred to as the *Alternative Hypothesis*, proposes that in the earliest stages of color evolution, color terminology does not partition the entire color space, but rather leaves some regions ambiguous and unnamed. Levinson's data collected on the Yélî Dnye people [76] lends support to this hypothesis. Accordingly, color terms initially remain focused around certain

substances or objects and these terms become more generalized over time, expanding further into the color space. Whenever the need arises for new color terms, they are introduced to fill in the unnamed regions. These gapped regions continue to get filled-in until eventually all gaps are minimized and the population is in agreement of color naming across the entire color space.

The above theories describe the evolution of categorizations in contrasting terms. As such, a number of literature has addressed the debate between these two hypotheses [79, 35, 69]. Given our unique approach to understanding the evolutionary nature of categorizations, we may be able to offer additional insights to this debate. While the main goal of this chapter is to develop and test our evolutionary framework, we take a few initial steps to evaluate these theories in the context of our evolutionary model (see Section 2.4).

Both theories make hypotheses about emergence of color categories, space partitioning, and color terms. Given the expansive lexicon of color terms contained in many languages, Berlin and Kay achieved a more targeted analysis through their formalization of basic color terms.

2.1.3 Basic Color Terms and the World Color Survey

The idea of "basic color terms" originates from the *universalist* perspective of the *linguistic relativity* debate. The universalist view was founded on the principle that color cognition is a universal, physiologically-based phenomenon. Since then, numerous studies have identified other factors influencing the evolution of color categories [16] (e.g. Jameson and D'Andrade's [51] interpoint-distance model showed that the 3D structure of perceptual color space is another factor that influences the partition of the color space). Berlin and Kay popularized the universalist view by exploring universal features of color categorization through a formalization of BCTs introduced in their book [10] (see Section 1.1).

The *World Color Survey* [66] which was undertaken in an effort to verify the claims made in Berlin and Kay's 1969 book [10] (see Section 1.2). The data from 108 of the languages included in this survey[2] serve as the basis for our new methodology to study the evolution of color categorization.

2.2 ColorSims: An Evolutionary Game Theoretic Tool

In this framework, naming strategies of simulated populations evolve to stationary equilibria as the agents play a simple communication game. A stationary equilibria is defined as a non-Nash stable system that undergoes minimal change over a long period of time. The communication game requires agents to assign names to color stimuli and "communicate" repeatedly with members of the population until a naming convention arises. The agents in these evolutionary dynamics are endowed with minimal perceptual and learning abilities. A Python-based program written by Sean Tauber, called *ColorSims* [131], was developed to model the communication game specified in Komarova et al. [73]. This model has been utilized in several past studies to evaluate the evolution of color naming systems in populations of simulated agents for a one-dimensional color space [73, 72, 58, 59, 60, 93].

The simulation framework used in this paper is an extension of *ColorSims* with (i) a higher dimensional color space which more realistically approximates human color perception, (ii) real observer population data taken from the WCS, and (iii) a measure to assess the stability of a color naming convention. This updated version of the software is referred to as *ColorSims 2.0* [38].

[2]Languages 62 (Mampruli) and 93 (Tarahumara-W) were omitted from analysis in this paper because of the poor nature of the data collection at the time of the survey. Any future mentions to "all WCS language" refers to the set excluding Languages 62 and 93.

2.2.1 Games

The evolutionary dynamics are comprised of two steps. First, agents discriminate between like and unlike stimuli by assigning names to the stimuli. Second, agents' naming strategies are either updated or learned from another agent's strategy depending on the agreement of their assigned names.

2.2.1.1 Discrimination-Similarity Game

The *discrimination-similarity game* was introduced by Komarova et al. [73] as a mechanism for populations of artificial agents to create shared categorization systems. In the game, agents make judgments about how "close" or "far" two stimuli seem from each other in their color appearance. In order for these judgments to be comparable across agents, they defined a measure *k-sim* which serves as an objective criterion for determining whether two colors are "close" or "far" in appearance. Komarova et al. [73] write,

> Page 363: $k_{sim}(a, b)$ is interpreted as being related to the utility of categorizing a and b as the same or different colors. It is defined by the environment and the life-styles of the individual agents. It is used to reflect the notion of the pragmatic color similarity of the patches. For instance, suppose one individual shows another a fruit and asks her to bring another fruit "of the same color." It is a nearly impossible task to bring a fruit of a color perceptually identical to the first, because different lighting, different color background and slight differences in fruits' ripeness contribute to differentiating its perceived color from the comparison fruit. Therefore to satisfy "of the same color" of a fruit's ripeness in practical terms, the individual must be able to ignore such unimportant perceptual differences and bring a fruit that is "of the same color" practically. It may also be as important to be able to distinguish ripe, edible, "red" fruit from the

unripe, "green" ones.

At the beginning of each round of play, two agents—Player 1 and Player 2—are randomly chosen from the population. Both players are presented with an independently and randomly selected pair of colored chips and are individually asked to provide a name for each chip. Players assign names to a chip based on their probability strategy matrix (see Section 2.2.4.3). Two chips should be considered of the same category and given the same color name if the two chips are within 1 *k-sim* of each other. Conversely, if their distance is greater than 1 *k-sim*, they should be considered of distinct categories and assigned different color names. A player is awarded a "personal success" if the assigned names of the two chips are consistent with the criteria above, and is awarded a "personal failure" if the criteria is not met. The round is considered a "social success" if and only if both players have personal successes and both assign the same names to the same chips, otherwise it is considered a"social failure". More specifically:

1. *Social success*: *Personal success* for Player 1, *Personal success* for Player 2 (e.g. chip a and chip b are outside 1 *k-sim*, Player 1 chose α for chip a and β for chip b, Player 2 chose α for chip a and β for chip b)

2. *Social failure*: *Personal failure* for Player 1, *Personal failure* for Player 2

3. *Social failure*: *Personal success* for Player x, *Personal failure* for Player y

4. *Social failure*: *Personal success* for Player 1, *Personal success* for Player 2 (e.g. chip a and chip b are outside 1 *k-sim*, Player 1 chose α for chip a and β for chip b, Player 2 chose θ for chip a and γ for chip b)

In the case of social success and social failure resulting from two personal failures—cases 1 and 2 respectively—players perform a form of reinforcement learning [73] by updating their probability matrices accordingly: in case 1, both players strengthen the probability

of choosing α and β when naming chips a and b respectively; in case 2, players weaken the probability of using their assigned names for chips a and b (see Section 2.2.1.3) This concludes a round of the *discrimination-similarity game*. For all other cases, players continue to another game, called the *2-player teacher game*.

2.2.1.2 2-player Teacher Game

The *2-player teacher game* was also introduced by Komarova et al. [73] as a particular implementation of a reinforcement learning dynamic. When the outcome of the *discrimination-similarity game* results in social failure but personal success for at least one player—cases 3 and 4 (see Section 2.2.1.1)—the same players from the previous game will further engage in the *2-player teacher game*. The game appoints the player with the personal success as the *teacher* (in the case of two personal successes, one will be chosen at random) and the remaining player as the learner.

The learner strengthens the probability of choosing the teacher's assigned names for chips a and b and weakens probability of choosing their assigned names for chips a and b. The teacher simply strengthens their probabilities for choosing α and β (see Section 2.2.1.3) This concludes a round of the *2-player teacher game*.

2.2.1.3 Reinforcement Learning

At the end of each round of the *discrimination-similarity* and *2-player teacher games*, the two players engage in some form of updating to their naming strategy matrices. These updates are described below and are arranged according to the four possible outcomes outlined in Section 2.2.1.1.

In the following descriptions, q represents the *reinforcement probability*. The *reinforcement*

probability is the unit by which an agent revises its belief about the assignment of a particular word to a particular color chip. In our simulations, $q = 1/(n * 4)$ where n is the number of words in the agents' vocabulary.

Also, $P^x_{i,\alpha}$ represents the probability that agent x will assign the name α to chip i. These values are revised by one *reinforcement probability* unit at a time according to the rules defined below. At all times, $P^x_{i,\alpha} \in [0, 1]$. If $P^x_{i,\alpha} - q < 0$ or $P^x_{i,\alpha} + q > 1$, then $P^x_{i,\alpha}$ remains at either 0 or 1 until $P^x_{i,\alpha} \pm q$ re-enters the range [0,1].

(i) Social success: Personal success for Player 1 and Player 2

Given that there is social success, it must follow that Player 1 and 2 assigned the same names, α and β, to chips i and j, respectively. Hence, Player 1 and Player 2 update as follows:

$$P^1_{i,\alpha} \rightarrow P^1_{i,\alpha} + q, P^1_{j,\beta} \rightarrow P^1_{j,\beta} + q$$

$$P^1_{i,\sigma} \rightarrow P^1_{i,\sigma} - q, P^1_{j,\tau} \rightarrow P^1_{j,\tau} - q$$

$$P^2_{i,\alpha} \rightarrow P^2_{i,\alpha} + q, P^2_{j,\beta} \rightarrow P^2_{j,\beta} + q$$

$$P^2_{i,\sigma} \rightarrow P^1_{i,\sigma} - q, P^2_{j,\tau} \rightarrow P^2_{j,\tau} - q$$

where σ and τ are two independently, randomly selected names other than α and β.

(ii) Social failure: Personal failure for Player 1 and Player 2

Suppose that Player 1 assigned the names α and β to chips i and j, respectively, and Player 2 assigned the names μ and ν to chips i and j, respectively. Player 1 updates as follows:

$$P^1_{i,\alpha} \rightarrow P^1_{i,\alpha} - q, P^1_{j,\beta} \rightarrow P^1_{j,\beta} - q$$

$$P^1_{i,\sigma} \rightarrow P^1_{i,\sigma} + q, P^1_{j,\tau} \rightarrow P^1_{j,\tau} + q$$

Similarly, Player 2 updates as follows:

$$P^2_{i,\mu} \to P^2_{i,\mu} - q, P^2_{j,\nu} \to P^2_{j,\nu} - q$$

$$P^2_{i,\sigma} \to P^2_{i,\sigma} + q, P^2_{j,\tau} \to P^2_{j,\tau} + q$$

where σ and τ are two independently, randomly selected names other than α and β.

(iii) Social failure: Personal success for Player x, Personal failure for Player y

Suppose that Player 1 had the personal success and assigned the names α and β to chips i and j, respectively, and Player 2 had the personal failure and assigned the names μ and ν to chips i and j, respectively. Then Player 1 is chosen to be the *teacher* and updates as follows:

$$P^1_{i,\alpha} \to P^1_{i,\alpha} + q, P^1_{j,\beta} \to P^1_{j,\beta} + q$$

$$P^1_{i,\sigma} \to P^1_{i,\sigma} - q, P^1_{j,\tau} \to P^1_{j,\tau} - q$$

Player 2 learns from Player 1's naming schema and updates as follows:

$$P^2_{i,\mu} \to P^2_{i,\mu} - q, P^2_{j,\nu} \to P^2_{j,\nu} - q$$

$$P^2_{i,\alpha} \to P^2_{i,\alpha} + q, P^2_{j,\beta} \to P^2_{j,\beta} + q$$

(iv) Social failure: Personal success for Player 1 and Player 2

Then a random player is chosen to be the *teacher* and the two players update their strategies as in case (iii).

The games are repeated until a stable, population-wide naming convention is reached.

2.2.2 Evolutionary Dynamics

The evolutionary dynamics arise through repeated rounds of the *discrimination-similarity game* and the *2-player teacher game*. The two agents and two stimuli used in each round are randomly selected before the round starts. Therefore, a simulation with r rounds and N agents will result in r/N interactions per agent on average. At the start of the simulation, social failures are high in frequency. As players engage in more interactions and update their naming strategies, a population-wide naming system begins to emerge as the level of agreement in chosen category names increases (see Figure 2.2). In other words, the population begins to partition the stimulus space similarly. If a population has reached an optimal, stable naming convention and continues to maintain that stability, the dynamics are considered to have reached a solution. In this study, a solution is considered to have reached convergence only when agent error is minimized and the optimal number of categories is reached. When optimality is obtained, all color categories are of roughly equal size. The solution to the evolutionary dynamics is always a non-Nash equilibrium: although error is minimized across the stimulus space, errors will persist at the category boundaries. This occurs because, at the boundaries, there always exist two chips, a and b, that are within *k-sim* that should both be named α but happen to belong to different categories and are consequently given different names. This kind of scenario continuously leads to personal and, therefore, social failures. Such errors do not influence the overall stability of the population-wide naming convention, though, as the certainty of names for non-border chips remains high. Hence, once the dynamics emerges on a solution, it is stochastically stable as the population-wide naming system experiences minimal change over a very long period of time (see Figure 2.3).

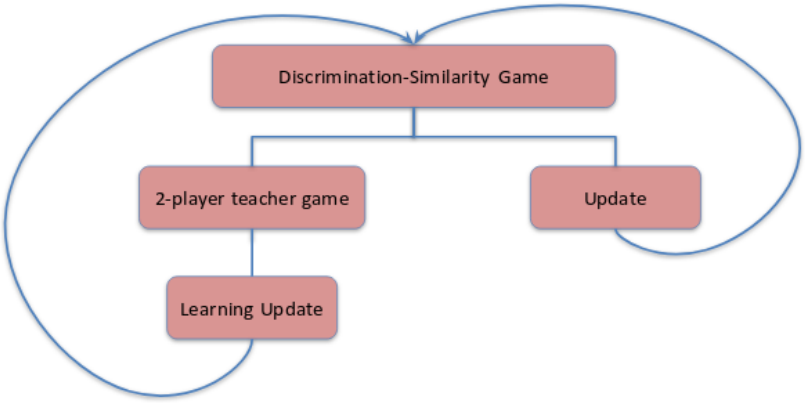

Figure 2.1: Flowchart of evolutionary dynamics.

2.2.3 Evaluating Solution Stability

There are two measures used to determine the stability of a population's color naming system:

1. The level of lexicon agreement across the whole agent population
2. The amount of change in the global agreement level across various time periods

Global lexicon agreement on round r of the game (A_r) is defined as follows [3]:

$$A_r = \sum_{i=1}^{320} \frac{P_{i,\alpha}}{N} \tag{2.1}$$

where r is the game round number, i is the chip number, α is the most frequently used name for chip i, N is population size, and $P_{i,\alpha}$ is number of agents who assign name α to chip i.

The global agreement level A_r is designed to be a measure of the strength of agreement between individual agents' color naming strategies within a population. A high A_r indicates high consensus in the population over their categorizations and low A_r indicates low

[3]For robustness, an alternate measure of global measure was defined called C_r. This alternate measure of agreement compares naming strategies between pairs of agents whereas A_r calculates agreement based by chip. These measures have been found to be highly correlated (r = 0.91). Therefore, for simplicity we use A_r exclusively. The definition of C_r can be found in Appendix A.

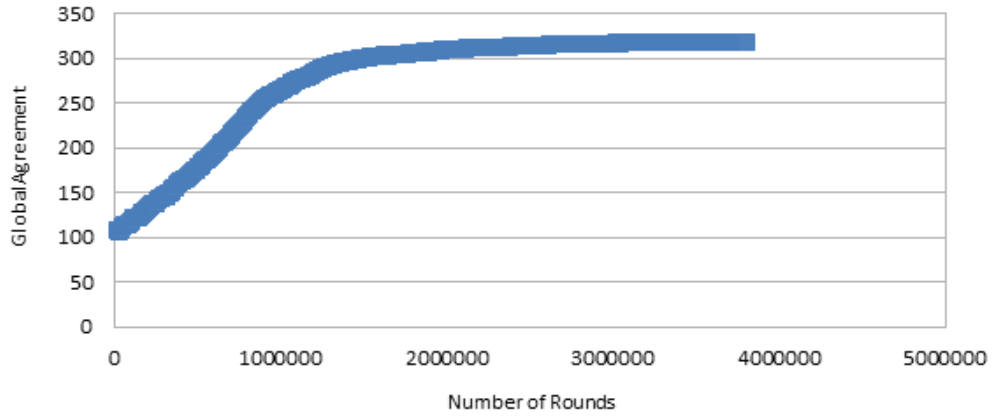

Figure 2.2: Plot of global agreement agreement measure (A_r) for a population of random agents across the number of rounds played.

consensus, or a large amount of variability between different agents' individual strategies.

The global agreement level is then used to calculate the *Stability Measure*, defined as follows:

$$S_r = \frac{A_r - A_{r-10,000}}{A_r + A_{r-10,000}} \tag{2.2}$$

Global agreement (A_r) is calculated every 10,000 rounds, so S_r is interpreted as the change in agreement level between rounds r and $(r - 10,000)$, or in other words, the change in agreement level between each "snapshot" of the population.

The simulation marks an r^*, which is the round number at which the stability measure first falls within some predetermined "stability range" (default range is (-0.00175, 0.00175), determined empirically) (see Figure 2.3). The simulation will stop if the solution stays stable—S_r stays within the stability range—for another r^* rounds. That is, a naming system is considered to have converged to a stable solution if the system is stable for r^* rounds, resulting in an overall total of $2r^*$ simulation rounds.

Defining S_r in this manner ensures two crucial properties: (i) the system will run for as long as it needs to reach convergence, as determined from S_r falling within a stability range, and

(ii) we can check that the solution remains stable over a sufficiently long period of time, as determined by r^*.

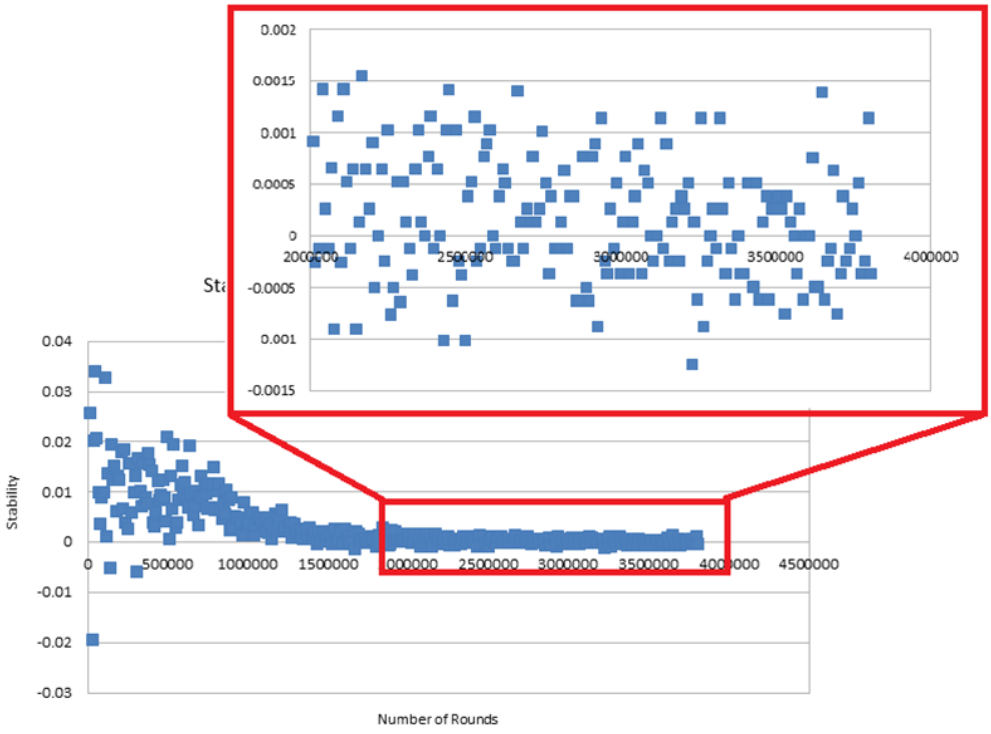

Figure 2.3: Plot of the stability measure (S_r) across numbers of rounds played. We define a "stable" solution to be one which falls within some "stability range" for many rounds. Due to the confusions that agents will have at the boundaries of two color words, the solution will be perpetually subject to minor fluctuations and changes.

2.2.4 ColorSims 2.0 Initialization

2.2.4.1 Higher-dimensional Color Space

The set of stimuli used in these simulations is a two-dimensional array of colored stimuli. Each individual stimulus in the array is referred to as a color chip. This array is a subset of the standardized set of 330 (320 chromatic, 10 achromatic) Munsell color chips, which is derived from a Mercator projection of a three-dimensional Munsell Book of Color perceptual

color space. The set of stimuli used in the simulations is identical to the chromatic component of the stimulus palette used in the World Color Survey (see Figure 1.2). Chips in the grid are organized along the rows according to eight Munsell lightness (Value) values and are organized along the columns according to the forty Munsell Hue values.

For the purposes of employing a standard color difference metric, each of the chips in this color grid is mapped to its corresponding coordinates in the three-dimensional CIELUV color space, which in essence allows for a principled assessment of agents' color discrimination judgments based on a standardized uniform perceptual color space. CIELUV was developed by the International Commission on Illumination (CIE) in 1976 with the aim of creating a perceptually uniform space. When presented with such stimuli in our evolutionary game, agents ostensibly "perceive the colors" as 3-tuples (L, u', v'), the coordinates of the colored stimuli in CIELUV space. While the set of stimuli is constricted to a two-dimensional grid, the underlying space that agents perceive and judge is represented as three-dimensional.

Whereas previous iterations of these simulations utilized a one-dimensional hue space, called a hue circle (i.e. a discrete array of color chips arranged according to just-noticeable-differences) [131], the simulation framework used in this study implements a three-dimensional perceptual color space in order to create a more realistic representation of how humans perceive color.

The CIELUV color model is appropriate for use here since its approximate perceptual metric permits computation of color difference, or Delta-E (ΔE), by employing the fact that Euclidean distances between colors map to similar distances in our perception[4]. Therefore, by using a perceptually uniform space, agents can use Euclidean distances between chips

[4]Both CIELAB and CIELUV were designed to be perceptually uniform color spaces. Therefore, in both spaces, ΔE can be computed using the standard Euclidean distance formula, which means either color space would suffice in our simulation model. For this procedure, we chose to use CIELUV instead of CIELAB because ΔE is a slightly better approximate for perceptual distances in CIELUV. However, we hypothesize that switching to CIELAB as the underlying 3D color space would not result in significantly different results. In the future, we will run simulations using different color spaces (such as CIELAB) to test the robustness of the results to changes in the assumed perceptual model.

in CIELUV as a basis for judgments of color discrimination when engaged in naming game scenarios used to evolve the population's naming conventions.

2.2.4.2 Color Discrimination Measure: *k-sim*

The discrimination-similarity measure *k-sim* is defined to be the minimum distance at which it becomes important, for pragmatic (and not perceptual) purposes, to treat two color chips as belonging to different color categories. Given the pragmatic importance of categorization, two chips that are within *k-sim* should be considered as belonging to the same category and two chips that are outside *k-sim* should be considered as belonging to different categories.

When the stimulus set is one-dimensional (i.e. hue circle), distance between color chips i and j is defined as the number of physical chips that are between i and j. However, when the stimulus set is two-dimensional, we define the distance between color chips to be their perceptual distance. Since the CIELUV space aims to be perceptually uniform, calculating color difference as distances in the physical space will in essence map to similar distances in human perception. Thus, distances between stimuli in the three-dimensional color space are calculated using the standard Euclidean distance metric:

$$\Delta E_{uv} = \sqrt{(L_1 - L_2)^2 + (u'_1 - u'_2)^2 + (v'_1 - v'_2)^2} \tag{2.3}$$

where ΔE_{uv} is the distance between two chips, (L_i, u'_i, v'_i) are the coordinates of chip i in CIELUV space.

For simulations using a one-dimensional color space, Komarova et al. [73] developed a mathematical formulation relating the optimal number of color categories to the total number

of chips in the stimulus set and the length of *k-sim* to

$$C^* = \frac{Q}{\sqrt{2k_{sim}(k_{sim}+1)}} \tag{2.4}$$

where C^* is the optimal number of terms and Q is the number of chips in the stimulus set. This formula can be used to derive the appropriate *k-sim* to use when initializing *ColorSims 2.0* simulations by setting $Q = 320$ and C^* equal to the number of BCTs identified by Berlin and Kay. Using these values, we can then solve equation (2.4) accordingly to find *k-sim*. When this formula was developed, the color space was assumed to be one-dimensional and consequently *k-sim* was defined as a distance according to just-noticeable-differences (*jnd*s). Though we are now utilizing multi-dimensional color spaces and measuring *k-sim* according to perceptual distances instead of *jnd*s, we justify using this method to find *k-sim* because, regardless of its underlying metric, *k-sim* continues to captures the same thing—how close or far colors are perceptually.

2.2.4.3 Agents

In the dynamics, every agent is endowed with a probability matrix that is updated at the end of each interaction. Agents are initialized with random probabilities as a basis to test the model and its parameters. Subsequent iterations of the simulation framework initialize agents with real observer data from the WCS.

Initializing Random Agents

Previous work using the *ColorSims* framework [73, 72, 58, 59, 60, 93] used populations of "random" agents. An agent's probability matrix is organized with colored stimuli along the columns and all possible color names along the rows. A random agent's matrix is populated with random fractions such that column j of the matrix defines a probability distribution for chip j over all possible color terms. Therefore, the value in cell (i, j) represents the

probability that this agent would assign the name i to color chip j. We employ the same process to initialize random agents in our *ColorSims 2.0* framework.

When an agent is asked to provide a name for chip j, the name that the agent assigns to chip j is the result of a probabilistic draw over the agent's vocabulary according to the distribution defined in column j of the probability matrix.

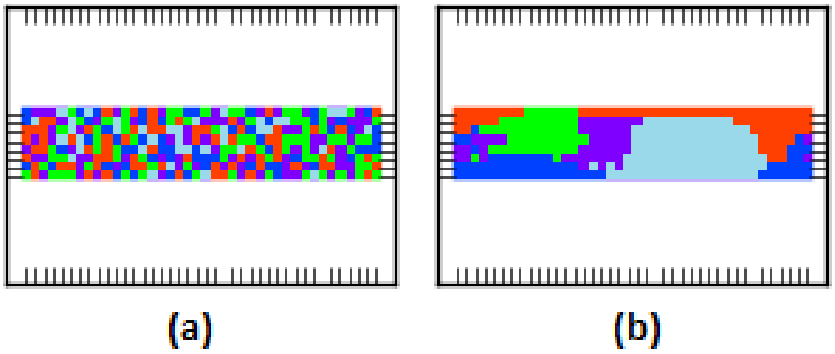

Figure 2.4: Visual representation of the color-naming system of a single agent in a random population at (a) the beginning of the simulation and (b) the end of the simulation. The rectangle at the center of each box represents the two-dimensional set of stimuli, where each pixel in the rectangle represents a single stimulus chip. The shade of each pixel represents the name in which the agent is most likely to assign to that stimulus.

Though these random agents are not representative of real observers, they are useful because they allow the evolution of the color naming systems to proceed *de novo*. By initializing random agents, we impose no restrictions on the initial state of the system categorization. Since agents can reach a shared, stable naming system with minimal assumptions, we can then evaluate how a system that is presumed to still be evolving, such as those from the WCS, might proceed to stability.

Initializing Agents with WCS Data

In our simulation-based investigation of color categorization, we initialize agent populations with the WCS naming-task data of each language. Since our simulations are limited to the 320 chromatic stimuli of the Munsell color grid, only the WCS data that was elicited from participants for those 320 chips is used in our analyses. For example, languages with 25

participants are represented by 25 agents. Each agent represents one of the WCS participants. A cell (i, j) in an agent's probability matrix is given a value of 1 if the participant named chip j with name i or was given a value of 0 otherwise (see Figure 2.5a). The agents' probability matrices are updated with repeated interaction between agents, characterized by the dynamics detailed in Section 2.2.1.

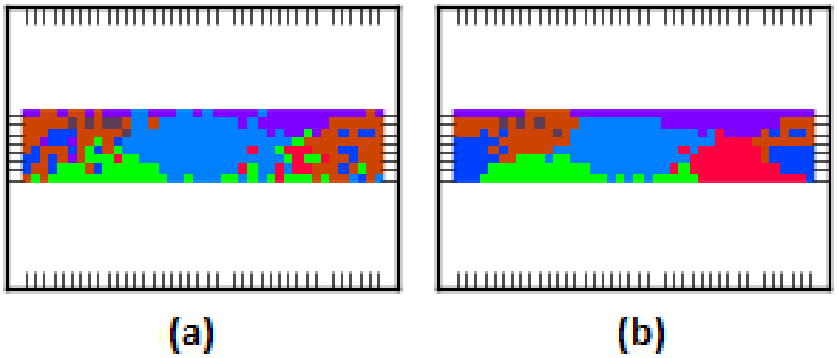

Figure 2.5: Visual representation of the color-naming system of a single participant from WCS Language 16 (Buglere) at the (a) beginning of the simulation and (b) end of the simulation. The naming strategy in (a) represents the actual names that this participant assigned to each of the chips in the rectangle at the time of the World Color Survey.

2.3 Results

Results are organized into two related but distinct sections: (i) a thorough investigation of the validity of our approach and (ii) a detailed analysis of the simulated data. To study the validity, we examine the simulation's influence on the original participant categorizations at both the population (global) and participant (local) level (see Section 2.3.1). The findings in the Simulation Influence analysis show that the end solutions maintained key features of the original categorizations and did not diminish the original integrity of the data. That is, the evolutionary dynamics did not impose an external structure on the original WCS naming systems. As such, we then developed methods to analyze the data which would provide insight into the categorizations' evolutionary processes.

2.3.1 Simulation Influence

2.3.1.1 Global Measure of Change

Over the course of the simulation, the global agreement of a population's categorization A_r is expected to increase as the agents update and learn through their interactions (see Figure 2.2). ΔA measures the amount of change from the original naming strategies of WCS participants to the converged solution based on A_r:

$$\Delta A = \frac{A_{\bar{r}} - A_0}{A_0} * 100 \tag{2.5}$$

Higher values of ΔA indicate that a greater proportion of chips changed names during the simulation. That is, populations with weak initial color naming systems (i.e. low A_0) required larger amounts of learning and chip reassignment to reach a shared solution. Conversely, lower values of ΔA imply that the population was already in high agreement with regard to its color naming convention (i.e. high A_0), and thus did not need much alteration.

The median value for ΔA across all WCS languages was 40%, the mean was 44%, and the standard deviation was 23%. The range of percent change extends from 8% to 123%. For 98.14 percent of all languages (106 out of the 108 languages used in this analysis), the percent change was found to be within one standard deviation from the mean. Additionally, for all WCS languages, $\Delta A > 0$, which indicates that the simulations were successful in improving the populations' naming system. Improvement is interpreted as the simulations removing previous ambiguity that existed in these populations' categorizations by allowing agents to learn and then develop a shared, common naming system.

The variability in the level of ΔA across the various languages of the WCS is unsurprising. There was no pragmatic reason for WCS participants to have high expertise across the

entire color space or to imply equal salience of all chips. Therefore, inter-group variation is to be expected given the improbability that most individuals would have color names for all 320 chips. Hence, the presence of this ambiguity when assigning names in the naming task necessitates that, at the minimum, some degree of learning must take place. Given the differing levels of initial population agreement, A_0, across all WCS languages, a variable degree of learning to achieve a stable solution took place.

While ΔA is useful in quantifying the total amount of change that a language undergoes during the simulation, it provides little information on how these changes impact each agent's underlying categorization. ΔA reveals what proportion of stimuli were reassigned to different names during the evolutionary process but provides little insight into whether these changes are a significant departure from each agent's initial categorization. Therefore, to determine whether the simulations preserve the integrity of the original WCS data, we analyzed the change in each participant's naming strategy.

2.3.1.2 Local Measure of Change

In the *mapping task* of the WCS, participants identified focal color(s) for a set of categories that had previously been elicited by the survey facilitators [66] (see Section 1.2). A *focal color*, or *best exemplar*, is a chip (or set of chips) that an individual participant identifies as the best example for a given color word. These chips are those that participants find as the most salient examples of a particular color category. The WCS database includes the complete set of focal colors that were identified by every participant in every language surveyed [23].

The focal colors of each participant were investigated as a means to understand the influence of the dynamics on the individuals' naming systems. We hypothesized that chips identified as salient by the observers are least likely to change categories during the simulated evolution.

By tracking the change to participant focal colors we analyzed the influence of the simulation, where low change indicates minimal influence.

A series of steps were completed to isolate the specific data needed for our analysis:

1. Focal colors that were located on the achromatic axis of the color grid were excluded since the achromatic axis was omitted from our simulations;

2. Only focal colors that were initially identified as focal for a BCT were considered for analysis;

3. Chips that were identified as focal for a term x but were assigned a different name y in the *naming task* were excluded from analysis due to the contradictory nature of these responses.[5]

This foci data was used to calculate proportion of each agent's foci that retained their original name. The average proportion across the population gives the measure of *focal color persistence*, formally defined as follows:

$$P_\ell = \frac{\sum_{i=1}^{N} \frac{|S_i|}{|F_i|}}{N} \tag{2.6}$$

where ℓ is the language number, N is the number of participants in language ℓ, F_i is the set of focal color terms for agent i, and $S_i \subseteq F_i$ is the set of focal color terms for i which have

[5]Because of the way BCTs were defined and WCS data was collected, the name assigned to a focal color need not be the name of the BCT for that chip. For example, if a population were to assign *crimson* as the name for a chip in the free listing task but *red* emerged from the analysis as the BCT for that chip, and it was determined that *crimson* was a subcategory of *red*, then *crimson* couldn't be a BCT, because, by definition, a BCT cannot be a subcategory of a larger color category. A more general consideration is that (i) given a category C and finding the chip a that is a focal of that category is a different task than (ii) given a and asking the participant to assign the best category name for a from a set of categories including C, even if the set only includes BCTs. This kind of asymmetry is discussed in detail in Jameson and Alvarado [56]. In our study the use of (ii) is the likely reason of why across languages a few foci were observed to switch names from free listing to BCT categorization. The WCS did not test for (i).

the same name at the beginning and end of the simulation. The range of P_ℓ is [0,1], where values close to 1 indicate high level of persistence within a language.

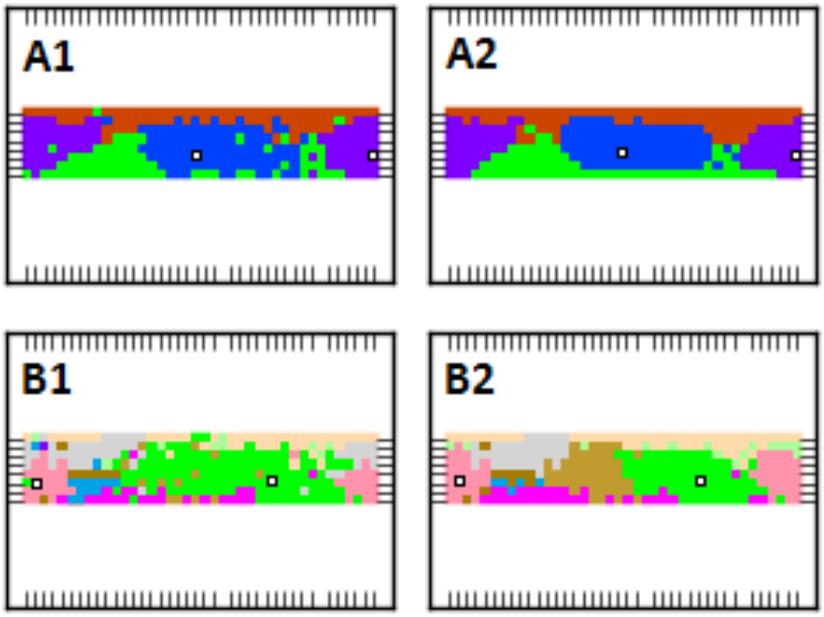

Figure 2.6: Naming systems for agents in Languages 74 (Múra Pirahã; A*) and 95 (Teribe; B*) at the beginning of the simulation (*1) and at the end of the simulation (*2) with markers for chips representing focal colors identified by those agents in the mapping task of the WCS. Foci are indicated using a white square with a black outline. Language 74 (A) is an example of a language with a small ΔA ($\Delta A_{74} = 0.19$). Language 95 (B) had larger ΔA ($\Delta A_{95} = 0.39$).

Based on our analysis, the median value for *focal color persistence* across all of the WCS languages was 0.97, the mean value was 0.94, and the standard deviation was 0.11. This result indicates that, on average, 94 percent of all focal colors identified for a language's basic color terms retained their original category name throughout the simulation (i.e. cognitively salient chips retain original names and learning takes place primarily on non-focal colors; see Figure 2.6).

Hence, despite the changes in each language's categorization during the simulation (evident from the percent change measures in Section 2.3.1.1), the change to the underlying structure of these naming systems was minimal. In other words, there is support that the communication mechanism used in these simulations did not imposing an external structure on the

original WCS data but instead improved the categorization across the populations while maintaining important features of agents' categorizations.

2.3.2 Simulation Analysis

To study the evolution of categorization, it is important to understand both the features of categories and what role these features play in the evolution. In this analysis, categories were defined as a collection of stimuli that share the same name. While chips located at a boundary of a category and at the center share the same name, there are important differences between the two (e.g. chips on the boundary are subject to persistent name changes while central chips are likely not). However, with no standard metric for identifying category boundaries, we developed a new probabilistic measure to identify chip location within a category. Knowing how each location of a category evolves helped us understand how a category is changing as a whole. Performing this location-based analysis provided insight as to what evolutionary processes are driving these observed category changes.

2.3.2.1 Identifying Boundary Chips

The WCS provides individualized data, per participant. This analysis is interested in population-level naming systems, including identifying population-level categories and category boundaries. Previous literature has proposed various metrics for identifying color category boundaries [30, 27, 15]. These methods utilized quantitative and statistical approaches in a static setting. Since our simulations used evolutionary dynamics and a pragmatic similarity measure, *k-sim*, these constraints must be incorporated into our measure of category boundaries. Hence, we developed a new probabilistic approach, based on *k-sim*, to identify category boundaries.

For any chip j in the 320-chip stimulus set, we first calculated the *boundary value* of chip j, which is the proportion of chips within 1 *k-sim* that have the same name as chip j.

$$B_{\ell,i}(c) = \frac{|S_c|}{|K_c|} \tag{2.7}$$

where ℓ is the language number, i is the numeric ID of the agent, c is the chip number, K_c is the set of chips within 1 *k-sim* of chip c, and $S_c \subseteq K_c$ is the set of chips within 1 *k-sim* of c that have the same name as c according to agent i. *Boundary value* increases as a chip becomes more central to a category—the higher the measure, the less likely that chip is a boundary chip.

Let the average *boundary value* for a given chip c across all participants in language ℓ be

$$\bar{B}_\ell(c) = \frac{\sum_{i=1}^{N} B_{\ell,i}(c)}{N}. \tag{2.8}$$

Then following method is employed to identify the likelihood that chip j is on the "boundary" of a color category. We define $f_\ell : [0,1] \to [0,1]$ to be the probability density function of *boundary values* for language ℓ such that $f_\ell(b) = P(\bar{B}_\ell = b)$, for $b \in [0,1]$. This function is constructed using aggregated frequency data from the *boundary values* of all 320 chips in the stimulus set.

Let $F_\ell : [0,1] \to [0,1]$ be the function that converts *boundary values* to *boundary probabilities* defined by

$$F_\ell(b) = \int_b^1 f_\ell(b)db \tag{2.9}$$

From this measure, we can then define the probability that a chip is part of a category "boundary" by

$$BP_\ell(c) = F_\ell(\bar{B}_\ell(c)) \tag{2.10}$$

The *boundary probability* measure, $BP_\ell(c)$ serves as an estimate for the topographic location within a color category. The higher the *boundary probability*, the more likely that a chip is on a category boundary. The lower the *boundary probability*, the more likely that a chip is central to a category.

2.3.2.2 Categorization Analysis

Another hypothesis was that some chips would evolve to a stable naming convention faster than others. Namely, central category chips, and possibly focal colors, due to their high salience, would reach full agreement more quickly than color category boundaries. Firstly, through the measure $BP_\ell(c)$, we calculated the probability that any given chip is on a category boundary. Secondly, we then tracked the chip agreement by calculating the proportion of the population that uses the most common name for each chip, remeasured at a fixed interval for the duration of the simulation. Chips reach full agreement when the proportion equals 1. Together these two measures allowed for a location-based analysis of categorization evolution.

To estimate *convergence rate* of a chip (i.e. the time for a population to reach full agreement on the chip name) we calculated mean of proportions,

$$CR_\ell(c) = \frac{\sum_{i=1}^{\bar{r}} P_{i,\alpha}(c)}{\bar{r}} \tag{2.11}$$

where $\bar{r}$ is the total number of game rounds, i is a round number, c is a chip number, α is the most frequently used name for chip c, and $P_{i,\alpha}(c)$ is proportion of agents in a population who assign name α to chip c. $P_{i,\alpha}(c)$ equals 1 when the whole population agrees on what name to assign to c. Therefore, the higher value of $CR_\ell(c)$, the sooner in the simulation that chip achieves total agreement.

To study the relationship between chip location within a category and the rate of convergence to full agreement, we calculated $corr(BP_\ell, CR_\ell)$ across all 320 color chips in the stimulus set. This provided an aggregated measure of how boundary probabilities relate to convergence time for each language.

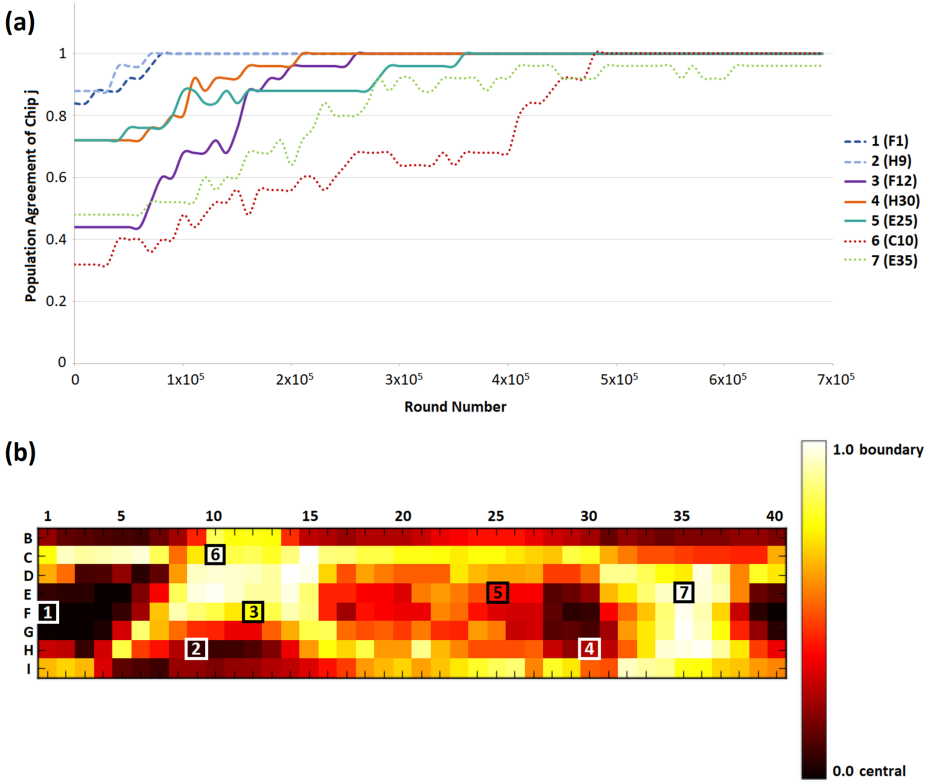

Figure 2.7: Simulations for WCS Language 23 (Chácobo): (a) Plot of chip naming agreements ($P_{i,\alpha}(c)$) for seven chips. (b) Heat map depicting the boundary probability ($BP_\ell(c)$) of each chip, organized on the 320 chip color grid. Boxed numbers on heat map correspond to the chips identified in the plot. From the results depicted in (b), we classify chips 1 and 2 as having *low* border probabilities, chips 3, 4, and 5 as having *moderate* border probabilities, and chips 6 and 7 as having *high* border probabilities. From (a), we observe that *low* boundary probability chips are the first to reach total agreement ($P_{i,\alpha}(c) = 1$), then the *moderate* boundary probability chips, and lastly the *high* boundary probability chips. This suggests that at a population level border chips are the last to be learned.

The correlations between BP_ℓ and CR_ℓ for all languages were negative, indicating that chips that are likely to be on a category boundary are also likely to converge slower (i.e. reach total agreement later in the simulation). Figure 2.7 presents a specific example of this finding.

Additionally, 88 percent of all languages had a correlation strength of |0.5| and above with close to 50 percent of languages at a strength of |0.7| and above (see Figure 2.8). Therefore, this measure of the relationship between boundary probability and convergence rate is non-trivial for a majority of the languages tested in this paper.

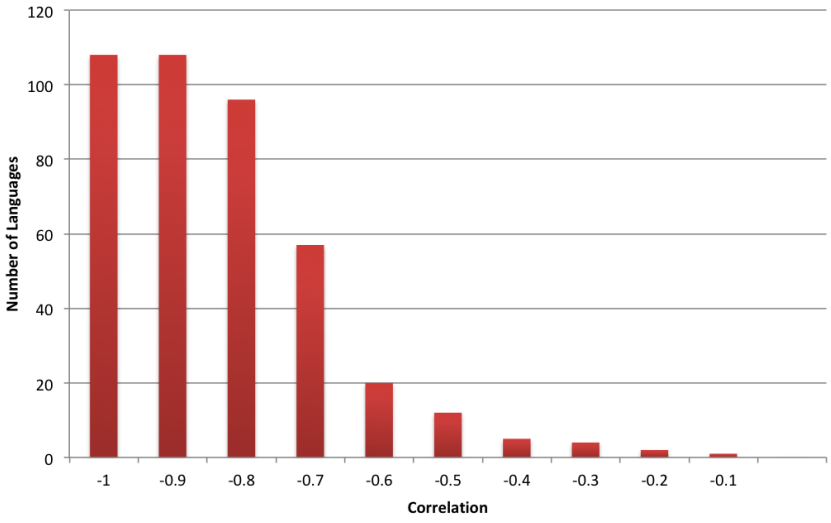

Figure 2.8: Relative frequency histogram of the correlations between the *boundary probability* and the *convergence rate* ($corr(BP_\ell, CR_\ell)$) of all WCS languages.

2.4 Discussion

The *Colorsims 2.0* framework allowed us to evolve each WCS language's categorization to a stable convention. While such an evolutionary process is not likely to occur in the real world, this approach can help us understand how categorizations can evolve under specific communication constraints. Importantly, it also allows researchers to test hypotheses within various theories such as those in *Partition Hypothesis* and the *Alternative Hypothesis*.

The methodologies detailed in *Simulation Influence* and *Simulation Analysis* (see Section 2.3.1 and 2.3.2, respectively) quantitatively describe the type of category evolution observed in the simulations. By analyzing the structure of the data in the end solution, we observed

that almost all foci maintained their original categorizations, despite variations in the amount of global system change. Though this may seem contradictory for languages with low initial agreement, it is reasonable to assume communities where color is not a salient concept have highly varied and sometimes inconsistent naming strategies, particularly for regions of color space the speakers have no pragmatic need to name. The *mapping task*, on the other hand, allowed participants to choose chips that best represented color categories, implicating that they likely chose chips which were highly salient. These salient chips thus served as "anchors" to the category and during the simulations, as the agents engaged in the learning dynamic, they gained expertise in these regions of confusion around the focal colors. Hence, the focal colors for a category persisted in the final naming system and learning took place at the category boundaries.

In the same vein, the analysis of the correlation between boundary probability measure, $BP_\ell(c)$, and chip convergence rate, $CR_\ell(c)$, revealed a substantial relationship for most languages. Therefore, there is evidence that chips located in the center of a category will reach full agreement more quickly than border chips. In other words, results indicated that the area of salience and expertise grows outwards from the focal colors to the ambiguous chips—a population gains expertise for central chips first and, with continuous communication, the expertise moves outwards towards the category boundary.

The above analyses showed that although each agent had a fully named color space, in aggregate, some WCS language populations exhibited regions with poor BCC agreement, where a single BCT was not agreed upon at the population level. By putting these population's categorization through an evolutionary dynamic, however, these regions with low intra-population consensus began to disappear. The results showed that the shrinking of these ambiguous regions was driven by the well-established, high consensus BCCs invading ambiguous regions until ambiguity is minimized. This process takes place while maintaining best exemplars of ℓ's BCCs throughout the dynamics. These results are reminiscent of the

Alternative Hypothesis of color evolution where categories begin with highly salient points and then extend outwards. In this way, *Alternative Hypothesis* is a better suited theory to explain the observed changes to the categorizations throughout the simulations. It should be noted though that the simulations kept *k-sim* fixed throughout and they were chosen in a manner so that the number of basic terms would likely be fixed. This handicapped partitioning occurring, because a category partitioning into two would require another category to disappear. Results of Komarova et al. [73] showed that *k-sim* varying with stimulus region but held constant for each region or its equivalent (i.e. having regions with different utilities) would not change the number of categories but change their sizes. Thus to change the number of categories regions *k-sim* must vary over time. This will be implemented in future work.

2.5 Conclusion

This study served as a preliminary exploration of the simulation-based approach, *ColorSims 2.0*, as a tool to analyze natural color categorizations of WCS communities that would otherwise be inaccessible to researchers due to missing empirical diachronic data. By applying communication constraints—features of communication that create a positive force towards the development of linguistic conventions depicted in *discrimination-similarity game* and *2-player teacher game*—on each WCS linguistic community, we were able to assess the level of stability of each language. This was achieved by using said constraints, along with minimal perceptual and memory assumptions for agents, to evolve each language to a state where a stable convention is shared across the entire population. Using the WCS as a test case was crucial to our ability to verify the validity of our approach. As hoped, the impact of these simulations on the original WCS data was minimal and our model imposed very little influence on the underlying structure of the categorizations. Yet, working with WCS data

carries its own limitations and impacted our analysis in the following two ways:

1. The WCS data provides a set of pre-existing categorization schemas with names for the entire set of stimuli—following Berlin and Kay's theory that BCCs perfectly partition the color space. With each population's fully labeled stimulus sets, Berlin and Kay chose the "best BCT" for each chip by taking the chip's modal color term. There, however, may exist little agreement about such a method, calling into question the usefulness of such "bests" in actual communicative practice. Similar concerns arise for the selection of a focal color for a BCT.[5] Despite such possible, implicit limitations in the WCS data, our approach extends the types of analyses that can be done using this data.

2. Theories of color categorization make claims about how categories evolve from their conception. While using the WCS data limited our understanding about how these categories evolved to that point, in later iterations we will update the evolutionary dynamics in such a way that better captures the evolutionary processes described by these theories. For example, for the *Alternative Hypothesis* we can initialize agent categorizations with a few salient points and allowing the other space to be learned while for the *Partition Hypothesis* start with WCS data and employ a mechanism to evolve *k-sim* to enable various category modifications (i.e. fission and insertion).

While these considerations may limit the extent to which theoretical implications may be drawn from these results, our preliminary study using WCS data in this evolutionary framework is promising and gives hope for further exploration. Future research will include developing a measure to examine the strength of a color categorization to provide a more detailed understanding of how categorizations are changing at the population level over time (i.e. semantic drift). Additionally, this measure will provide the basis for a new criterion to group populations for inter-language analysis, similar to Bimler's use of the quantitative in-

dex of inter-language similarity to determine groupings of WCS languages [15]. While there have been methods developed to categorize naming systems according to stages of evolution, these methods lack information about how established a language is within its evolutionary stage. Performing this analysis using simulation-generated data would afford us the means and opportunity to draw specific conclusions about where a language is in its evolutionary process.

Chapter 3

Evolving Unnamed Space in Color Naming Systems

3.1 Introduction

The debate of nature versus nurture has been discussed in many fields and across a wide variety of contexts. In language and cognition, this debate is centered around the concept of linguistic relativity. Linguistic relativity is concerned with the influence, if any, that language has on cognition or thought. This question of if and how language affects cognition is one that has been studied in disciplines such as anthropology, philosophy, linguistics, and cognitive science. Some scholars believed that language did influence thought (nurture) while others held that human cognition and perception were driven by universal, biological factors (nature). This debate was popularized in the field of color naming with the release of Brent Berlin and Paul Kay's book *Basic Color Terms: Their Universality and Evolution* [10], sparking two opposing views: the *universalist view* and the *relativist view* [110].

The universalist view argues that color naming is governed by universal properties which are

shared among all humans, such as the physiology of our visual system. According to this view, color naming systems are subject to universal constraints and therefore must be drawn from a bounded set of states. Though this view was made popular through *Basic Color Terms* [10], arguments in favor of universal constraints date back to the mid-nineteenth century—the most notable work being that of William Gladstone [36] and Lazarus Geiger [33]. Both Gladstone and Geiger studied ancient written languages and noticed that different patterns of speech among speakers of these ancient languages in comparison to speakers of modern European languages [110]. They noted that color naming was more inconsistent and less precise among these ancient languages, causing them to suggest an evolutionary process for color lexicons which mirrored the biological evolution of color sensation [110]. At the turn of the century, however, relativism became the prevailing mode of thought [105, 106, 37]; that is, until the introduction of *Basic Color Terms* [10].

Berlin and Kay revived the universalist perspective by positing that the set of possible color words was universally constrained and that languages acquired these terms in a systematic order as they evolved in complexity. They argued that the observed variation in color naming systems from different languages was due to them being at different stages of evolution. Field research of the Dani language in Papau New Guinea performed by Eleanor Rosch Heider [44, 43] also lent support for the universalist view. Heider conducted color naming and memory tasks among both English and Dani samples and found that focal colors in both languages consistently had the shortest names and the fastest recall [43]. The observed salience around focal colors for a language with few basic color terms (Dani has 2) and one with many basic color terms (English has 11) was presented as evidence in favor of the universalist view. Berlin and Kay, in an effort to find empirical support for their hypotheses, collected the *World Color Survey* (WCS; see Section 1.2). Analyses concluded that the findings from the WCS [66] data were consistent with their original universal ideas.

Conversely, the relativist view asserts that color naming is influenced by culture-specific

phenomena and inter-language variation is caused by unique cultural features. In their review of the color naming debate, Regier et al. state, "The relativist view...holds that semantic distinctions are determined primarily by largely arbitrary linguistic convention" [110], meaning that there are no restrictions or constraints on how a lexicon might partition the color space. This view is closely related to the idea of linguistic relativity and the Sapir-Whorf Hypothesis. The Sapir-Whorf Hypothesis asserts that a speaker's thoughts and actions are determined by their native language [68]. This principle can be expressed in a strong form—language determines thought—or in a weak form—language influences thought.

The relativist argument grew in prominence throughout the beginning of the twentieth century as anthropologists and anthropological linguists began studying foreign languages which exhibited systematic features, comparable to that of European languages [110]. Some of the relativist literature from the mid-twentieth century preceded *Basic Color Terms* [105, 106, 37], but several were written as a challenge against the ideas of by Berlin and Kay [83, 82, 117]. Recent support for the relativist view can be attributed to the work of Debi Roberson and colleagues who have conducted field work on the Berinmo language in Papau New Guinea [114, 112, 113]. Roberson et al. found that the color category boundaries of Berinmo did not align well with those of English. Particularly, they noted that the Berinmo had a sharp boundary between the yellow and green regions of the color space that arose for practical reasons, not from perceptual processing constraints. The category boundary serves the following purpose: "tulip leaves, a favorite vegetable [among the Berinmo], are bright green when freshly picked and good to eat, but quickly yellow if kept. Agreement over the color term boundary coincides with agreement over when they are no longer good to eat and is highly salient in a community that talks little about color." [112]. Therefore, since this boundary served an important pragmatic purpose, salience across this boundary was much higher for the Berinmo speakers than it was for English speakers [114]. Roberson used the example of the Berinmo as a counterexample to the universalist claims.

For centuries, academics have passionately advocated for their side of the debate. The prevailing theory of the time has oscillated between the universalism and relativism multiple times. Regier et al. describes this movement as a swinging pendulum [110]. Several articles have been written in direct response to claims made by proponents of the opposite camp [82, 117, 53, 54, 109]. One of the main points of disagreement in this long-standing debate is the process by which color naming systems evolve over time. The universalist theory of evolution is called the *Partition Hypothesis* while the relativist theory is called the *Emergence Hypothesis*.

3.1.1 Partition Hypothesis

The *partition principle* is an idea that was present in the original Berlin and Kay model [10] and its subsequent reformulations [9, 70, 67, 65], but wasn't formalized until 1999. Kay and Maffi defined the *partition principle* to be the notion that languages will jointly partition the perceptual color space using a limited set of basic color terms (BCTs) [69].

> Partition acts minimally and incrementally. We begin with the color space lexically partitioned into just two cells, that is, named categories, each cell (named category) representing a union of some subset of the six fuzzy sets corresponding to the primary colors and then at each new stage reapplication of partition and the other three principles [discussed below] adds a single new cell (i.e., term), until the six primaries have each received a distinct basic color term. (Kay & Maffi, 1999: 750)

The *Partition Hypothesis* (PH) describes the evolutionary process under the assumption of this principle. Since the PH requires the entire color domain to be named, color categorizations from preindustrial societies are assumed to evolve from simple, high-consensus systems

to more complex, high-consensus systems [79]. When evolution occurs, one of two things can happen: (i) either an existing category will split to become two new categories (e.g. the splitting of "grue" into "green" and "blue") or (ii) a new category will insert itself between two existing categories (e.g. the insertion of "turquoise" between the "green" and "blue").

Evolution under the PH is built on three color appearance-based principles [69]:

1. Black and White (Bk&W): Distinguish black and white
2. Warm and Cool (Wa&C): Distinguish warm primary colors (red and yellow) from cool primary colors (green and blue)
3. Red (R): Distinguish red

These three principles in addition to the *partition principle* are applied in the following order to determine the color terms present in any given stage of evolution:

$$Partition > Bk\&W > Wa\&C > R$$

For example, at Stage I, a system is assumed to have two color terms. By *partition*, the two terms must jointly partition the whole of the color space. The Bk&W principle necessitates that black and white be encapsulated in separate terms. Wa&C necessitates that the term containing red and yellow should be separate from the term containing green and blue. For reasons further discussed in [69], warm colors are associated with white and cool colors are associated with black. Thus, a Stage I language will have two BCTs: white/red/yellow and black/green/blue. Applying the same principles in Stage II results in the three term system: white, red/yellow, and black/green/blue. The same process is repeated for Stages III–V, yielding the nine systems depicted in Figure 1.1 in Section 1.1. Systems progress through the stages as their culture becomes more technologically advanced and acquires new color words, following a path designated by the arrows in Figure 1.1. The PH is defined by this

"refinement" of existing partitions (e.g. splitting of composite categories; white/red/yellow → white & red/yellow) when the need for a new term arises.

3.1.2 Emergence Hypothesis

The name *Emergence Hypothesis* (EH) was first dubbed by Kay and Maffi [69] to address the work of Levinson [76], Lyons [85], Lucy [83, 82], and Saunders and van Brakel [117] which rejected the *partition principle*. The EH, unlike the PH, does not assume that all languages possess a set of color words which completely partition the color space [69]. Drawing on empirical evidence gathered from languages which do not fit the universalist model [76], the EH claims that BCTs first emerge through salient, physical objects. Since the EH posits that color words refer to objects in a speaker's immediate surroundings, it is reasonable to assume that there might be large unnamed regions of the color space due to the sparse occurrence of certain colors in nature. If an individual never encountered a certain color, there would not be a pragmatic need to assign a name to that color. Though the EH asserts that color terms first emerge through a small set of salient objects, it does allow for the emergence of a separable semantic domain (similar one assumed under the PH) as the color categories expand over time [69, 76].

One of the most referenced empirical exceptions to the PH and main evidences in favor of the EH is the Yele language spoken by the Yélî Dyne tribe from Rossel Island. The language was studied by Steven Levinson who found that it exhibited features which were incompatible with Berlin and Kay's model of BCTs [76]. In Yele, there are three color terms which meet the criterion for basicness: *kpaapîkpaaî* (white), *kpêdê* (black), and *mtyemtye* or *taataa* (red). The presence of the terms black, white, and red as BCTs is consistent with Berlin and Kay hypothesis regarding the order in which color terms are acquired [10]. However, the Yélî Dyne's color naming system violates the assumption that the set of BCTs

will jointly partition the entire color domain. The words in Yele which are translated to English's white, black, and red are not color terms as we understand them—referring to a class of many discriminable colors—but are, rather, object specific. The word for white, *kpaapîkpaaî*, is derived from the word for a species of cockatoo with pure white feathers. Terms for red, *mtyemtye* or *taataa*, are in reference to a crimson-colored parrot. Finally, the term for black, *kpêdê*, is the word for a tree which has brown bark and produces nuts that turn black once they have been roasted in a fire to make them edible [76]. Because of the specificity of these color names, the Yele categories for white, black, and red are not large composite categories as are expected of a typical Stage II language. This counterexample to the Berlin and Kay model caused Kay and Maffi to amend the model to accommodate the existence of languages which were inconsistent with its original claims [69].

3.1.3 Recent Developments in Color Naming

Following several reformulations to Berlin and Kay's evolutionary encoding sequence [9, 70, 67, 65], Paul Kay and Luisa Maffi released a revised evolutionary theory which attempted to reconcile the universalist framework with Levinson's findings on the Yélî Dyne [69]. To accommodate non-partition languages, such as Yele, Kay and Maffi included a new stage which preceded the BCT encoding sequence. They asserted that there are two kinds of evolution: evolution *towards* a basic color term system and the evolution *of* basic color term systems [69]. Thus non-partition languages are hypothesized to be in the process of evolving towards a fully partitioned state and will subsequently begin the process of "refining their partition" just the same as partition languages.

While Kay and Maffi did define a model which accounted for some of the relativist counterexamples, their model still had a distinct universalist flavor to it. Since then, many researchers have abandoned the two opposing viewpoints by acknowledging that both universal and

culturally-relative factors affect color naming. Some have even gone so far as to argue that "the oppositional framing itself should be jettisoned altogether, since it has outlived its usefulness and is an obstacle to understanding" [110]. Recent findings clearly demonstrate that there are universal constraints on color naming and that differences in color naming across languages do cause differences in color cognition [71]. Among these findings are several studies which still acknowledge universalities among cross-language color naming systems but reject the *partition principle*.

Lindsey et al. collected empirical color naming data from the Hadza people—a group of nomadic hunter-gatherers in Tanzania—and performed information-theoretic analyses to investigate properties of languages with very simple lexicons [79]. The color naming task prompted informants to provide color names to a set of 23 Munsell color chips. Unlike previous studies that utilized a similar naming task, subjects in this study were allowed to answer with "don't know." Lindsey and colleagues concluded that Hadzane, the language spoken by the Hadza, was in very early stages of color category evolution. Black, white, and red were the only terms used with high consensus, and for non-black/white/red samples "don't know" was used more frequently than any other term. By performing information-theoretic analyses on the data, the authors were able to show that the Hadza fell on the same continuum as the World Color Survey languages. Therefore, they noted that participants from WCS languages which were near Hadzane on this continuum likely would have used "don't know" more often if they had been permitted to [79]. Thus, despite the fact that the WCS data was collected under the assumption that all colors must be named, they showed that several WCS languages exhibited features similar to that of Hadzane. This finding challenged the assumption that the set of BCTs must jointly partition the whole color space [69].

Zaslavsky et al. demonstrated that empirical color categorizations can efficiently compress meanings into words (i.e. perceived colors into color terms) by optimizing the information bottleneck trade-off between the accuracy and complexity of its lexicon [141]. The

information-theoretic model that they built treated color categories as soft categories with weighted membership, not hard partitions of the color space [10, 69]. Since membership to a category was not discrete, areas of low consensus could easily be identified and visualized. A simulated annealing along the information bottleneck curve which bounded the convex frontier of possible naming systems revealed a projected evolutionary sequence. This sequence predicted of the order in which color terms would enter a system and was found to match closely with Berlin and Kay's encoding sequence [10]. In addition to revealing this hierarchy of categories, the evolutionary sequence showed where these categories might emerge. They found that new categories entered the system in regions of low consensus, as others had previously theorized [86, 76]. This finding was an important step in modeling the continuous evolution of color categorizations and understanding the transitional phases between evolutionary stages.

Clair Hill studied the color naming systems of speakers of an Aboriginal Australian language called Umpila and found an interesting connection between the gaps in their color categorizations and kin relationships [45]. Umpila is a three color term system with high consensus in the terms black, white, and red. Similar to the Yélî Dyne [76], Umpila speakers left large regions of the color space unnamed. Though the Umpila informants could not provide color words for many of the color samples presented to them, they were still able to make sense of the unnamed samples and relate them to the named samples using their grandmother-grandchild kin relation. For example, the speakers would pair the "red" samples as grandmother with "like red" (pink-ish in hue) samples as grandchild. Thus, Umpila speakers are able to conceptualize the gaps in their color naming system by drawing on knowledge from other domains, such as kin relationships. Not only does this finding provide further empirical evidence for languages which fail the *partition principle*, but it also explores the forces, even those which come from other domains, that influence and shape the organization and understanding of these color naming systems.

3.1.4 Rationale For This Model

The evolutionary framework proposed in this chapter builds on the model introduced by Komarova et al. [73]. They found that shared color categorization systems could emerge among a population of agents through a simple learning and communication mechanism and without any perceptual constraints. Several other models have complexified features of this framework to be more representative of reality, such as the implementation of heterogeneous observer populations [72, 58, 59, 60], agent memory [93, 128], changing populations under a birth-death dynamic [102], and the initialization of agents using real observer data [39].

Building on the foundation of previous evolutionary models, this study seeks to explore the evolutionary mechanisms of color naming systems when specific domain knowledge is absent (i.e. when there are gaps in a language's color naming systems). Of the many models that have built off the Komarova et al. model [73], none have relaxed the assumption that all colors must be named. The motivation for including empty, or unnamed, space in the agents' naming systems comes from empirical evidence showing that multiple pre-industrial languages [76, 79, 45] exhibit these kinds of gaps. In color naming surveys, it is reasonable to suppose that a participant might respond with "don't know" when asked to provide a name for a color they are unfamiliar with [79]. If asked to guess a name for an unknown color, they might search their color naming system for the nearest named color and then guess that color's name. This notion is consistent with the physical property that color categories are convex and that "near" colors should be given the same name [73].

The proposed evolutionary model mimics the above scenario. Every agent is first endowed with a set of centroids—these act as anchors to the agent's color categories. The centroids could represent salient objects which are tied to the language's color words, such as the white cockatoo, red parrot, and tree that produces black-ish nuts in Yele [76]. When an agent encounters a color c that he is unfamiliar with, he guesses a name for c by considering

the distance of c to each centroid. The closer c is to a centroid c^*, the more likely the agent is to use the name attached to c^* for c. Thus, the probability that an unnamed color will receive a term is determined by its distance from other more salient colors.

By initializing agents with gaps in their color naming systems, we can evaluate how these gaps evolve and change as a result of repeated interaction. Modeling agents in this way is useful for better understanding the early stages of color category evolution. Recent studies have empirically demonstrated that pre-industrial languages with simple lexicons exhibit regions of unnamed space [76, 79, 45]. Therefore, the model described in this chapter reflects features of early stage languages and approximates the mechanisms present at the beginning of color lexicon evolution. Since our methods are based on evolutionary dynamics, we can both characterize features of this empty space and evaluate the processes by which the empty space is filled.

3.2 Methods

The evolutionary dynamics proposed in this study are an amendment to the simulations presented in Chapter 2 [39]. Some of the simulations performed in Chapter 2 initialized agents using *naming task* data from the World Color Survey. In the *naming task*, participants were required to provide a color name for all of the presented stimuli. Therefore, by using forced naming task data, the evolutionary model from Chapter 2 was developed under the assumption that the entire color space needed to be named. However, in light of several studies demonstrating the empirical presence of unnamed regions of the color space, it seems prudent to allow for gaps in the agents' color naming systems. This model implements many of the mechanisms included in previous evolutionary game theoretic models [73, 39] and presents new features to model unnamed space.

3.2.1 Discrimination Similarity and 2-Player Teacher Games

The evolutionary algorithm utilized here is modeled after the *discrimination similarity game* and the *2-player teacher game* first introduced by Komarova et al. [73]. The *discrimination similarity game* represents the process by which pairs of agents are presented with a pair of stimuli, assign a name to each stimuli, and then are judged on whether their assignments were reasonable given the objective truth *k-sim*. The discrimination-similarity measure *k-sim* is defined to be the minimum distance at which it becomes important, for pragmatic (and not perceptual) purposes, to treat two color chips as belonging to different color categories [73]. This game can result in either personal success or personal failure for each agent and either social success or social failure for the pair collectively. The result from the *discrimination similarity game* determines the updating action that will be performed in the *2-player teacher game*. The *2-player teacher game* is similar to simple reinforcement learning, but the agents take on the roles of teacher and student in certain outcomes. An agent becomes a teacher when it has a personal success and its partner has a personal failure. In this case, the failed agent learns from the teacher by adopting the teacher's naming pattern (see Section 2.2.1.2 for detailed description of these updating rules). Komarova et al. [73] were able to demonstrate that this learning dynamic alone was sufficient for a population of agents to develop a shared categorization system.

The model developed for this study also utilizes the *discrimination similarity* and *2-player teacher* games, and adds to the existing line of research by allowing the agents to leave regions of the color space unnamed.

3.2.2 Evolutionary Model

Empirical, diachronic data on color categorizations is extremely sparse and very difficult to collect. Using simulation-based approaches to study color category evolution is advantageous

because it broadly approximates evolutionary trends, generating idealized diachronic data to fill the empirical void [39]. Also, many scholars develop theories of color category evolution, assuming particular mechanisms are in place (e.g. how a new term is acquired or how a name gets associated with a category) without specifying the mechanics behind the mechanism [127]. Therefore, defining explicit computational models is helpful for understanding and evaluating the factors which impact evolution. This study follows the same goal of previous simulation-based approaches to color naming; specifically, seeking to investigate how evolutionary mechanisms are affected when non-salient regions of the perceptual color space are introduced. Elements of this model include the physical color space from which stimuli are drawn, the individual agents who are grouped into populations, and the dynamics which dictate how the agents communicate with and learn from each other.

3.2.2.1 Color Space

The color space used in this model is the two-dimensional grid of Munsell color chips used in the WCS (see Figure 1.2 in Section 1.2). As in Chapter 2, each of the individual color chips is mapped to its (L, u', v') coordinates in CIELUV color space in order to calculate the distance between them, ΔE (see Equation 2.3 in Section 2.2.4.2). The 10 achromatic chips (A0–J0) are omitted from analysis, leaving the set of 320 chromatic color chips to be set of stimuli used by the agents.

3.2.2.2 Agent Initialization

Agents in this framework are modeled to represent individual participants from the WCS. A collection of agents is a population. The individual agents are modeled using the participants' *naming task* data. Populations represent all participants from a single WCS language. The initial structure of the agent's color naming systems is determined by a "filling" parameter,

denoted by θ. The parameter θ ranges between 0 and 1 and represents the proportion of 320 chromatic Munsell color chips which will be named at the outset of the simulation. Intuitively, the lower θ is, the more unnamed space the agent starts out with and the higher θ is, the more of the original participant's data is used to fill the agent's naming strategy[1] (see Figure 3.1).

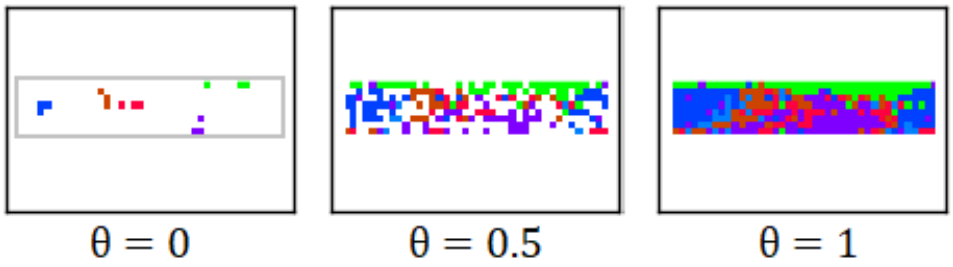

Figure 3.1: Initialized agent naming systems with varying levels of θ. The colored rectangles represent the Munsell color grid and the color of each pixel in the rectangle denotes the name the agent is assigning to the stimulus. The colors are used for visualization purposes only and do not represent the hue of the underlying color space. White regions represent unnamed stimuli. θ represents the probability that non-focal chips will be named at time $t = 0$.

Each agent is endowed with a vocabulary of size n, where n is the number of unique terms used by all participants in the WCS language, and a probability matrix which represents the agent's naming strategy. The matrix has dimension $n \times 320$. The value in cell (i, j) of the naming strategy represents the probability that the agent will name chip j with name i. In previous iterations of this framework [73, 39], the j-th column of the naming strategy was required to sum to 1 at all times, defining a probability density function over the set of possible names for chip j. In this iteration, however, the columns of the naming strategy need not sum to 1. The removal of the restriction that $\sum_{i=1}^{n} p_{i,j} = 1 \ \forall j = 1, ..., 320$ is how unnamed space is introduced into the system.

Initializing an agent's naming strategy follows the following process:

[1] A simulation with $\theta = 1$ is identical to the simulations described in Chapter 2.

1. Let n_c be the number of centroids to identify for each color category[2].

2. Identify n_c centroids for each of the Language ℓ's BCTs: centroids are determined by calculating the *boundary value* (see Equation 2.7 in Section 2.3.2.1) for each chip named with the BCT. The n_c chips with the highest *boundary values* are chosen as the centroids.

3. For each centroid, c, find the cell (α, c) in the agent's naming strategy corresponding to the chip c representing the centroid for the color term α and set the value in the cell equal to 1.

4. For the chips in the stimulus set that are not centroids, define a Bernoulli random variable X such that the outcome $X = 1$ represents the chip being given a name and $X = 0$ represents the chip staying unnamed. The $P(X = 1) = \theta$ and consequently $P(X = 0) = 1 - \theta$.

5. Take a draw from X for each chip. If $X = 1$, set cell (α_j, j) equal to 1, where α_j is the color term which the original participant assigned to chip j. If $X = 0$, let $\sum_{i=1}^{n} p_{i,j} = 0$ for chip j.

When prompted for a name for a color chip, an agent consults his naming strategy. Since the columns of the naming strategy are not required to sum to 1, an agent will find himself in one of three cases—each of which he will take a slightly different approach to picking a name (see Figure 3.2).

Case 1: $\sum_{i=1}^{n} p_{i,j} = 1$

The agent already has a fully defined probability distribution over the set of available terms.

[2]In the *mapping task* of the WCS, participants were asked to provide a stimulus or set of stimuli which were best exemplars for a given color term. That is, best exemplars were not limited to a single color and could be represented using a set of colors. We draw on this intuition when initializing the centroids, allowing $n_c \geq 1$.

In this case, the agent takes a probabilistic draw over the set of terms according to $p_{i,j}$ $\forall i \in \{1, ..., n\}$.

Case 2: $\sum_{i=1}^{n} p_{i,j} = 0$

The agent has no idea what name to assign to chip j. In an empirical setting, this is a scenario in which the agent would respond with "don't know" if asked to name chip j. Therefore, he will use his knowledge of the initialized centroids to make a guess. He takes a probabilistic draw where the probability of picking category c is inversely proportional to chip j's distance (in CIELUV space) to category c's centroids. To transform distances in perceptual color space to a measure of similarity, the distance between the two color chips is passed through the function $f(d) = e^{-0.0003d^2}$.

Case 3: $0 < \sum_{i=1}^{n} p_{i,j} < 1$

The agent has some existing notions about what names to assign to chip j but also retains some uncertainty about what name to use. In this case, for the names i such that $p_{i,j} \neq 0$, the agent picks name i with probability $p_{i,j}$ and with probability $1 - \sum_{i=1}^{n} p_{i,j}$, the agent responds that he is unsure of what to name j. When the agent doesn't know what name to assign, he takes a probabilistic draw according to the distance of j from the category centroids, as described in Case 2.

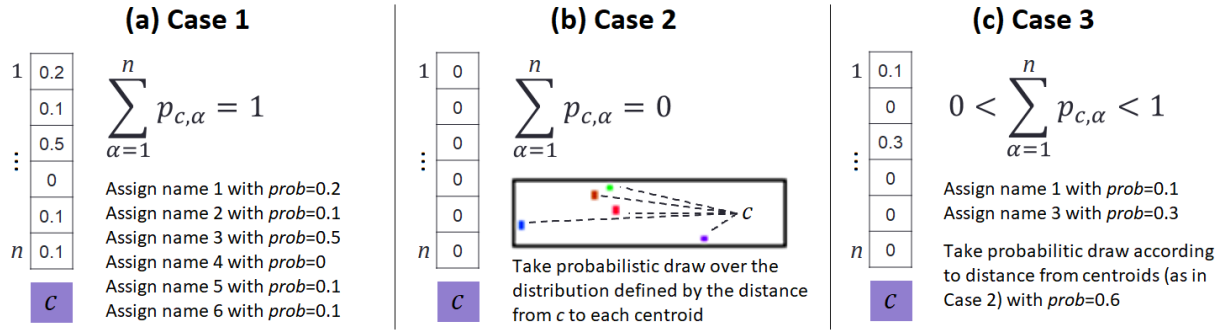

Figure 3.2: Illustrated example depicting how agents assign names to color chips. The column populated with probabilities represents the column of the probability matrix corresponding to color chip c. The value in each cell represents the probability with which the agent will assign name α to c.

Agents are organized on a complete network, meaning that every agent is connected to, or able to communicate with, every other agent in the population.

3.2.2.3 Evolutionary Dynamics

The dynamics of this model are facilitated through the implementation of the *discrimination-similarity* and *2-player teacher games* (see Section 3.2.1). Each round of the simulation, agents engage in pairwise interactions according to the following process:

1. Two agents are randomly drawn from the population.
2. Two color chips are randomly sampled from the color space.
3. Agents consult their naming strategies in order to provide a name for each stimulus.
4. Agents individually engage in the *discrimination-similarity game* to determine whether the terms they used were appropriate given the objective reality *k-sim*.
5. The outcome of the *discrimination-similarity game* for each agent (i.e. *personal success* or *personal failure*) determines what rules the agents will use to update their naming strategies and whether they will move on to play the *2-player teacher game* (see Section 2.2.1.3).

This process is repeated until the system is considered to have reached a stable solution.

3.2.2.4 Convergence

In order to ensure that the solutions obtained by the simulations are stable solutions, a stability criterion is defined which determines when a simulation should be stopped. This stability criterion is based on the *Stability Measure* (see Equation 2.2) described in Section

2.2.3. The *Stability Measure* is calculated every 10,000 rounds and represents the amount of learning that occurred between rounds $r - 10,000$ and r. Once the measure reaches some ϵ-neighborhood around 0 and remains in this neighborhood for a "sufficiently long"[3] time, we consider the system to have converged to a stable solution (see Figure 3.3).

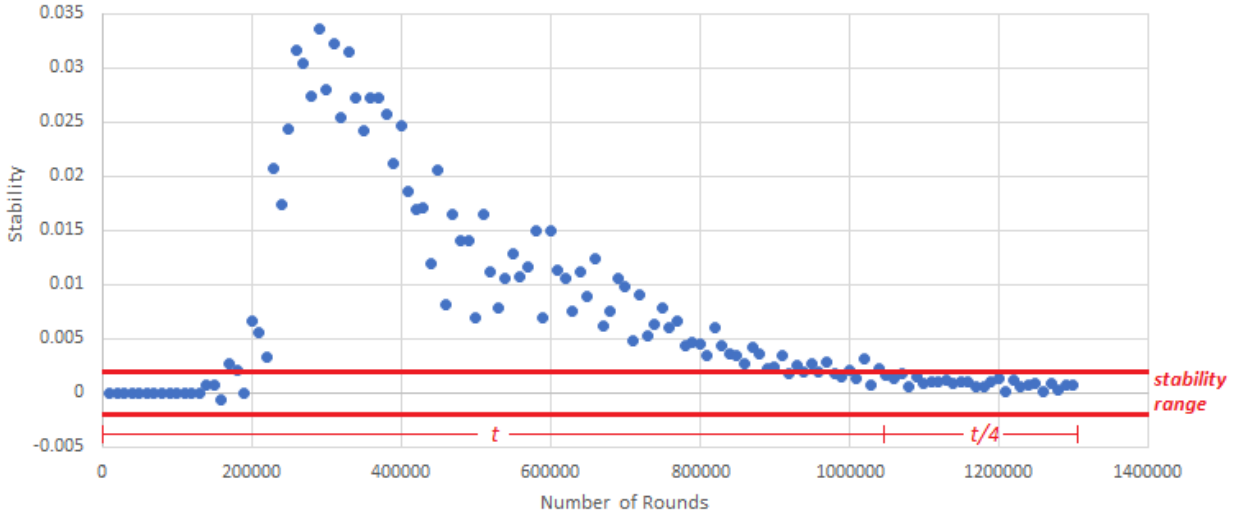

Figure 3.3: Plot of the Stability Measure over time depicting the system reaching a converged state. Once the system enters the stability range, it stays in the stability range for $(1/4) * t$, where t = number of rounds it took to enter the stability range.

As evident in Figure 3.3, the *Stability Measure* starts and stays at zero for many rounds. This is due to the way the agents' naming systems are initialized. At the beginning of each simulation, the agents are instantiated with very strong notions about what term to assign to each color chip. (The rationale for this approach is rooted in the fact that the agents are informed using empirical data that was collected from real observers.) Therefore, to give the population ample time to begin the learning process, a fixed number of rounds is set to occur before the model begins checking for convergence.

[3] "Sufficiently long" is determined to be (1/4)*number of rounds it took to enter the ϵ-neighborhood. In Chapter 2, convergence is reached when the stability measure stays inside the stability range for as long as it took to enter the stability range. This criterion is loosened for this study because the intuition behind the stability range was to identify when change became minimal. Since change is assumed to be minimal in the stability range, it is sufficient to stop the simulation closer to when the stability range is first entered.

3.2.3 Model Implementation

A subset of the 110 languages collected in the World Color Survey was chosen to be evaluated under the described evolutionary dynamics. The 25 languages studied are described in Table 3.1, including the language name, country of origin, sample size, number of BCTs, and proposed evolutionary stage (the stages listed here correspond to the stages depicted in Figure 1.1). This subset was picked to be representative of the whole survey; that is, the distribution of the number of BCTs across languages in the subset is proportional to the distribution of the entire survey. For each language, a series of eleven simulations were initialized using $\theta = \{0, 0.1, 0.2, ..., 1\}$.

The simulations were executed using an amended version of ColorSims 2.0, a Python package designed to simulate the evolution of color naming systems [38]. ColorSims was altered in order to accommodate the introduction of unnamed space into the agents' naming systems. The changes to the Python scripts reflect the processes and principles described in Section 3.2.2.

Each simulation was run for a variable number of rounds, determined by the time it took for the *Stability Measure* to enter and stay in the stability range. After empirical testing, the stability range was set to be $(-0.0002, 0.0002)$ and the number of fixed rounds to play before checking for convergence was 400,000 rounds. The number of centroids, n_c, initialized for each BCT for each agent was set to 3.

3.2.4 Comparing Naming Systems

Following the completion of a simulation, we want to compare the resulting agent naming systems within and also between different populations. To achieve this, inter-naming system

Lang #	Language	Country	# Speakers	BCTs	B&K Stage
14	Bété	Ivory Coast	25	3	II
104	Wobé	Ivory Coast	25	3	II
18	Campa	Peru	25	3	II $\rightarrow$ III$_{\mathrm{G/Bu}}$
66	Mayoruna	Peru	25	4	III$_{\mathrm{G/Bu}}$
74	Múra Pirahã	Brazil	25	4	III$_{\mathrm{G/Bu}}$
64	Martu-Wangka	Australia	25	4	$\rightarrow$ IV$_{\mathrm{Bk/Bu}}$
1	Abidji	Ivory Coast	25	5	II $\rightarrow$ III$_{\mathrm{G/Bu}}$
47	Iduna	Papau N. Guinea	25	5	IV$_{\mathrm{G/Bu}}$
59	Kuna	Panama	25	5	IV$_{\mathrm{G/Bu}}$
89	Sursurunga	Papau N. Guinea	26	5	IV$_{\mathrm{G/Bu}}$
101	Vasavi	India	25	5	IV$_{\mathrm{G/Bu}}$
83	Podopa	Papau N. Guinea	14	6	III $\rightarrow$ IV$_{\mathrm{G/Bu}}$
10	Arabela	Peru	25	6	$\rightarrow$ IV$_{\mathrm{G/Bu}}$
9	Apinayé	Brazil	30	6	IV$_{\mathrm{G/Bu}}$*
52	Kamano-Kafe	Papau N. Guinea	25	6	$\rightarrow$ V
70	Micmac	Canada	25	6	V
25	Chayahuita	Peru	25	7	IV$_{\mathrm{G/Bu}}$
5	Amarakaeri	Peru	6	7	IV$_{\mathrm{G/Bu}}$ $\rightarrow$ V
39	Guahibo	Colombia	25	7	IV$_{\mathrm{G/Bu}}$ $\rightarrow$ V
80	O'odham	USA/Mexico	25	8	IV$_{\mathrm{G/Bu}}$
27	Chiquitano	Bolivia	25	8	V
19	Camsa	Colombia	25	9	V
38	Garífuna	Guatemala	28	10	V
106	Yakan	Philippines	25	11	V
17	Cakchiquel	Guatemala	30	12	V

Table 3.1: The subset of 25 WCS languages used in the simulations.

comparisons are quantified using the Rand Index [104]:

$$R_{i,j} = \frac{|S_{i,j}| + |D_{i,j}|}{\binom{320}{2}} \tag{3.1}$$

where $S_{i,j}$ is the set of pairs of chips for which agent i uses the same name for both chips and agent j uses the same name for both chips, $D_{i,j}$ is the set of pairs of chips for which agent i and agent j both use different names for each chip, and the denominator represents the total number of pairs (color grid is comprised of 320 color chips). Since the Rand Index is only concerned with whether two color samples belong to the same group or not, it is able

to create an linguistically-agnostic measure, permitting cross-language comparisons.

3.2.5 Naming System Visualization

As described in Section 3.2.2.2, agent naming strategies are defined by a probability matrix, where each cell contains the probability that the agent will assign a color name to a color chip. To visualize the agent's partition of the color space, these probability matrices are transformed into word maps. A word map is an 8×40 matrix representing the set of 320 chromatic Munsell color chips. The value in each cell reflects the name that the agent is most likely to assign to that chip. This is usually done by taking the maximum value of each column in the probability matrix (i.e. finding the color name with the highest probability of being used). However, in order to facilitate the presence of unnamed space, a cell in the word map is only populated if there is at least one color term with a greater than 50% probability of being used to name that color. That is, an agent will only manifest a name for a color if he is fairly certain about which name to use. In visualizations of an agent's word map, each cell (i.e. color chip) is colored according to the name the agent is assigning to that chip. If the agent has not yet learned a name for a color chip, in the visualization the chip will be uncolored/white (e.g. see Figure 3.1). Word maps are used to visualize the agents' naming systems over the course of the simulations and are also used to perform certain comparative analyses.

3.2.6 Types of Unnamed Space

Unnamed or empty space represents colors which informants would respond with "don't know" when prompted for a name. There are several reasons why such a response might be elicited. One possibility is that the informant has neither encountered that color before nor has pragmatic need to name it, resulting in the absence of any associated color word for

the specified color [79, 76]. Another reason could be that there exist many competing words in the population's lexicon for that color, giving rise to an ambiguity about which name to assign to the specified color [79].

To differentiate between these two types of unnamed space, we consider the consensus among the population's individual word maps. For each of the 320 colored stimuli, we aggregate the color category names (including "don't know") used by all agents in the population. Unnamed colors can be further distinguished. If there is at least one name which more than 50% of the population is using for that color, then the color is considered to have a salient name among the population. Otherwise, the color is considered unnamed. If the modal answer among agents for a certain color is "don't know", then that color is considered part of *no name* space. Alternatively, if a majority of the agents are assigning a name to that color but the names being assigned are diverse, then the color is said to be in an *ambiguously named* region of the color space. Tracking the consensus of names over many rounds of the simulation will provide insight into the evolutionary mechanics which work to fill in this empty space and generate more complete population naming systems.

3.3 Results

From a total of 275 simulation runs (25 languages each initialized with 11 values of θ), all were able to reach a stable, shared naming convention. This convergence can be observed in Figure 3.4, as all simulation runs depicted had a learning rate which converged to a minimum and remained minimal for a significant amount of time. Similar to the results from the evolutionary model implemented in Chapter 2, populations had increased levels of agreement (see Equation 2.1 in Section 2.2.3) in their converged solution. Final average agreement among agents in a population, $\bar{A}_r = 0.908$, was found to be much higher than the initial average agreement, $\bar{A}_0 = 0.373$. Also, the converged systems had more contiguous

categories than the initial systems (see Figure 3.5). These two findings demonstrate that the communication mechanism is successful in providing agents with clearer discriminating abilities as well as higher agreement about the meaning of each color category.

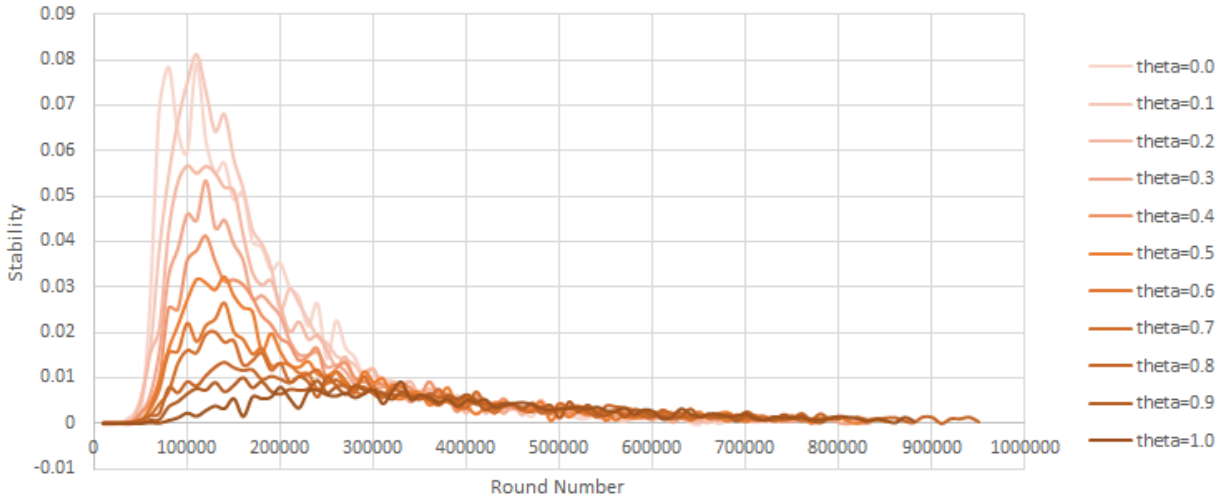

Figure 3.4: Population learning rates for simulations with varying levels of θ. Learning rates are represented by the *Stability Measure*.

Another noteworthy result is that all of the populations, regardless of their value of θ, reached near full partition status by the end of their simulation. 98.17% of color chips in all end solutions had a category name, and the small percentage that did not were usually located on a category boundary, where there were two or more competing color words. Figure 3.5 depicts the evolving naming systems of two agents, one with $\theta = 0$ and another with $\theta = 1$, both initialized using data from Participant 1 of Language 10 (Arabela). Both agents, despite having very different initial systems, were able to achieve a full partition of the color space.

One of the main questions that this study seeks to address is how the introduction of unnamed space into the initial naming systems affects the evolution of the color lexicons. In order to evaluate the effects of unnamed space, we first examine the structure of the end solutions across values of θ. The following process is used to compare the end solutions with different levels θs:

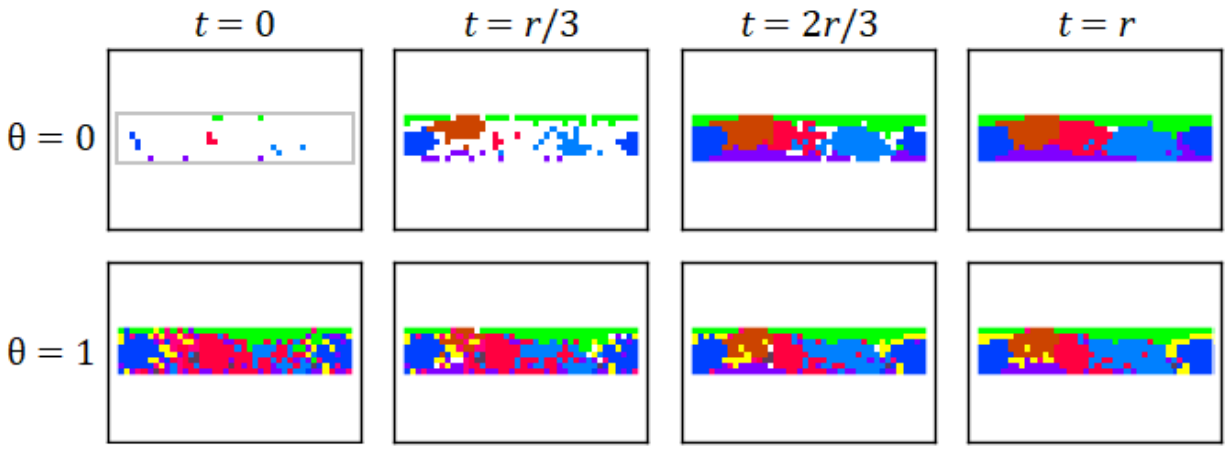

Figure 3.5: Naming systems of the same individual from Language 10 (Arabela) but initialized with different θs over the course of the simulation. The number of rounds in the simulation is represented by r. The colored rectangle represents the Munsell color chips and the color of each pixel represents the name assigned to that chip. Colors used are for visualization purposes only and do not reflect the hue of the underlying space. Chips that are colored white have no assigned name yet.

1. Consider the 11 populations, $\theta = \{0, 0.1, ..., 1\}$, representing WCS Language ℓ.

2. For each population, compute the modal map of the end solution by determining the name used with the highest frequency for each of the 320 color chip stimulus set.

3. Compute the Rand Index, R_{θ_i,θ_j} between the two modal maps for every pairwise combination of populations.

4. Let the average Rand Index of Language ℓ be $R_\ell = \frac{1}{\binom{11}{2}} * \sum_{\theta_i,\theta_j} R_{\theta_i,\theta_j}$.

The value R_ℓ was computed for every language included in Table 3.1. The resulting average Rand Index across all 25 languages in the sample was $\bar{R}_\ell = 0.913$ and the standard deviation was $\sigma_{R_\ell} = 0.042$. Since a Rand Index of 1 represents perfect agreement between two systems, the high $\bar{R}_\ell$ observed indicates that the final naming systems of populations modeling the same language but with different values for θ are highly similar in structure (see Figure 3.5 at $t = r$).

Population consensus over individual color chips was tracked over the course of the simulations in order to evaluate evolution from a local perspective. At time t, population consensus was considered for each stimulus to determine whether the color has *no name*, is *ambiguously named*, or is *saliently named* (see Section 3.2.6). Colors which fell into *no name* space was coded as a -2 for analysis, colors in *ambiguously named* space were coded as -1, and colors that were *saliently named* were coded as 0. Therefore, for each stimulus, we can define a ternary vector which tracks the classification of a color (using the values -2, -1, and 0) across the rounds of the simulation (see Figure 3.6). The correlation between the classification vector and a vector containing the number of game rounds at each simulation snapshot was computed to determine if there was a systematic pattern in the changes in classification of a color over time. This correlation was computed for every language and resulted in an average correlation of $\bar{\rho} = 0.668$ with a standard deviation of $\sigma_\rho = 0.152$. Positive correlation indicates that as time progresses, colors move from one of the unnamed stages, *no name* or *ambiguous*, towards being *named*. Colors appear to generally follow the following trajectory:

1. A majority of the population starts off with no color word attached to a color.
2. Agents will individually begin trying out different words, resulting in a diverse population naming system.
3. After repeated communication, a high-consensus word eventually arises and a stable convention is established.

Additionally, the location of a chip within its category was examined to see if there are any topological effects or dependencies on the evolution of the chip. To determine the location of a chip, we computed the *boundary probability* of the chip (see Equation 2.10 in Section 2.3.2.1), which is a measure that increases as a chip gets closer to a category boundary. To quantify the learning trajectory of each chip, we identified the round at which a chip becomes transitions from being *unnamed* to *named*. The correlation between the boundary

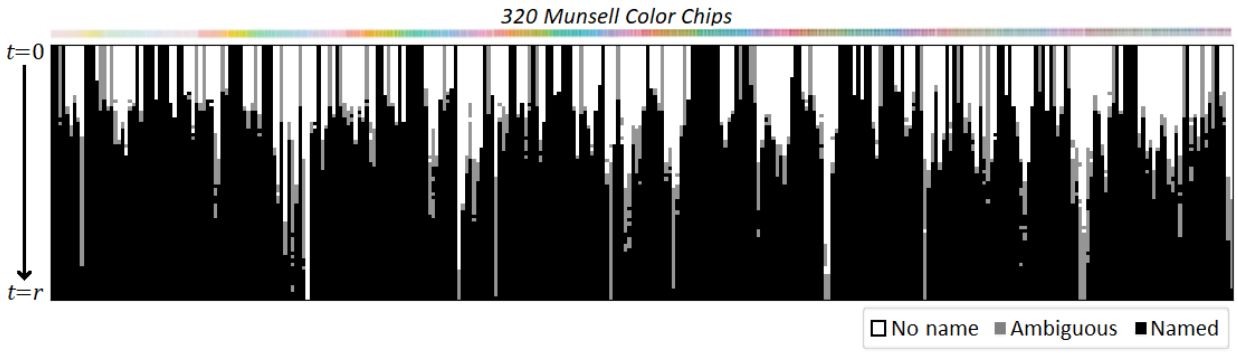

Figure 3.6: Trajectory of color chips from unnamed to named for simulation of Language 106 (Yakan) with $\theta = 0.5$. Each column of the visualized matrix represents a color chip. The rows represent the simulation rounds. Color chips generally transition from white to gray to black as time increases.

location of each chip and the round number at which it becomes *named* is computed for every language. The average correlation across all languages was $\bar{\rho} = 0.668$ with a standard deviation of $\sigma_\rho = 0.152$, indicating that location does affect the learning path of a color. Since the correlation is positive, this means that colors located near the boundary take longer to become named and thus spend more time in the *unnamed* stages. This suggests that learning starts at the center of categories and expands outward towards the boundaries in a systematic fashion (see Figure 3.7). Though others have posited this same phenomenon [39, 76, 79], the advantage of our and results is that we explicitly examine the state of the population's naming system as it undergoes evolution. For example, we observe this kind of "ripple effect" where chips transition from having *no name* to being *ambiguously named* to being *saliently named* and these transitions first happens among categorically central chips then move outward systematically towards the category boundaries.

3.4 Discussion

Previous work utilizing evolutionary game theoretic models to study color category evolution have demonstrated the sufficiency of a simple learning and communication mechanism to

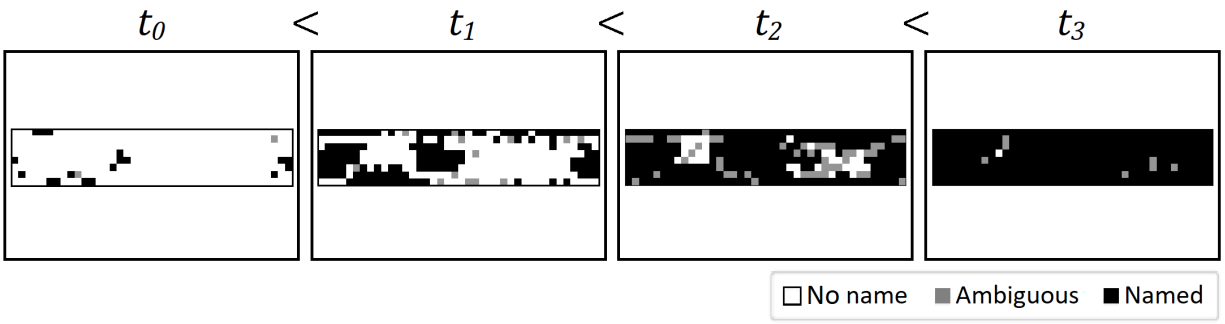

Figure 3.7: Mapping of unnamed space over the course of a simulation of Language 74 (Múra Pirahã) with $\theta = 0.3$. Learning appears to expand from the center of categories and expand outward to fill empty space.

develop shared naming conventions [73, 39]. The results presented in this chapter further validate those claims and show that populations can converge on a stable naming conventions even when agents are initialized with sparse information. That is, simple learning was able to not only organize messy information into meaningful categorized systems, but also to fill in gaps in understanding by drawing on information contained in a few key anchors. This result is also consistent with claims made by existing literature that new color terms will start with initially low consensus and will gradually evolve to a state of high consensus [79]. Thus, the *partition principle* is learned through repeated interaction and communication, not maintained throughout the course of evolution. Lindsey et al. write of the Hadza, "For individual informants, the representation of color was incomplete, but collectively, Hadza usage of color terms showed the beginnings of a more complete color naming system" [79]. Just as the Hadza were hypothesized to be on the path to a more complete naming system, we found that populations of agents that started out with incomplete color naming systems could achieve a fully named system through repeated interaction and communication. This finding supports the claim that even diverse and sparse naming systems can be on the path towards a complete, partitioned naming system.

The system solutions obtained through the evolutionary framework were not only complete, but they also retained structural similarity to the data used to initialize the agents. The

few key anchors (i.e. initialized centroids) proved to be sufficient to generate systems that were reminiscent of those which were were evolved using complete domain knowledge (i.e. systems with $\theta = 1$). The average Rand Index across the different pairs of θ reveals that the all of the evolved systems from a single WCS language are structurally similar to each other. In Chapter 2, we showed that our evolved color naming systems still maintained key features, namely placement of the focal colors, of the original data. Since the names of the focals persisted throughout the evolutionary process, we argued that the evolved naming systems were capturing the underlying structure of the original WCS data. The simulations performed in Chapter 2 are represented in this chapter by simulations with $\theta = 1$. Therefore, the high value of the Rand Index between simulations with different levels of θ indicates that sparse systems (i.e. those with low levels of θ) can evolve into a complete system which captures structural similarities to the empirical reality reflected in the WCS data.

As mentioned in Section 3.1.3, others have addressed the presence of unnamed space in color naming systems, and a few have used quantitative methods to do so [79, 141]. However, the claims made by these studies were limited by the use of static methods[4]. By studying these regions of low salience under an evolutionary framework, we were able to explicitly characterize the dynamic nature of this space as categories expand outwards from their centroids. We found that agents had multiple motivations for responding with "don't know" when prompted for the name of a color: (i) either the population had no available term for the color or (ii) there were too many competing terms to establish a clear convention. We also found that the same color could be considered unnamed for one of these two reasons at different points along the simulation. Colors appeared to follow the trajectory of having *no name*, to being *ambiguously named* due to the presence of a diverse set of competing names, to eventually establishing a high-consensus term. This finding suggests that color naming conventions are established by agents first communicating and teaching their salient

[4]Zaslavsky et al. [141] were able to generate an evolutionary trajectory by annealing their information-bottleneck curve. However, their information theoretic methods were static.

terms to others and then by learning others' known terms until a shared, stable convention is established for every stimulus in the color domain. Since these populations were modeled using features that have been examined in languages at early stages of evolution, these results suggest the dynamic by which early categorizations evolve towards a fully partitioned system. Performing this evolutionary analysis on static data provides explicit insight into the dynamics of the population and its lexicon that has not been afforded by previously applied methods.

The evolution modeled in this chapter is fundamentally different from the evolutionary theory proposed by Berlin and Kay [10, 69]. While Berlin and Kay define a discrete set of stages with varying complexity and characterized evolution as movement through these stages, this model treats evolution as a continuous phenomenon. Others have also modeled color category evolution in a continuous form, both theoretically and computationally [86, 76, 79, 141]. Rather than fitting an evolutionary trajectory through a set of pre-defined points, the implementation of a continuous model allows classifications, such as stages, to be extracted from the "naturally" occurring trajectory. For example, we classified colors into three different types of space by evaluating consensus at different times throughout the simulation. These classifications arose as an artifact of the methodology (e.g. by considering population-wide consensus over time) instead of purely from theory, and were thus observed as a result of features of the simulated populations. An advantage of using computational models to study color category evolution is this ability to define evolution continuously, as it is in reality, to better understand the properties and dynamics of these transitional stages.

3.5 Conclusion

The aim of this study was to examine the evolution of color naming systems in a linguistic population when regions of the color space are left unnamed. Modeling unnamed space in

the agents' naming systems is an important extension of previous evolutionary models of color categorization [39] in light of recent empirical work which highlights the existence of unnamed space in pre-industrial languages [76, 79, 45]. We found the addition of unnamed space results in converged naming systems which are similar in structure, regardless of how much unnamed space is introduced. Also, we were able to more closely examine the properties of this unnamed space and its evolution, building on the findings of previous researchers with regard to the presence of this unnamed space [79, 141]. This study further demonstrates the usefulness of evolutionary models to examine aspects of evolving conventions which would have otherwise been extremely difficult if not impossible to study in an empirical setting.

Though these evolutionary models are helpful in studying specific features of evolution, often they make certain assumptions which prevent them from accurately representing the hypothesized evolutionary processes. For example, the model described in this chapter assumes a fixed number of color terms in the agent's lexicons. In almost all theoretical models of color category evolution, as systems become complex, they acquire new color terms. However, in this model, the number of foundational color terms is established at $t = 0$ with the initialization of the category centroids. This initialization is based on data from languages which are not necessarily at an early stage of evolution and thus is not an accurate representation of the underlying processes. A more representative model would start with a small number of color terms and would systematically introduce new terms as the system grows in complexity [141]. This future work could provide keen insight into better understanding the mechanisms at play in the evolution of empirical color naming systems.

Another assumption made by this model that could be addressed in future iterations is the way systems are initialized according to the filling parameter θ. Here, non-centroid chips are randomly given a name according to the probability θ when, in actuality, named colors are more likely to be near the centroids. That is, empty space is not randomly distributed across the non-centroid colors, but is distance dependent based on the categories' anchors.

Making this change would most likely strengthen the results found regarding the transition of colors from unknown to named, as the regions of study would be more contiguous instead of randomly dispersed.

The results obtained from this study provide useful insight into the evolutionary dynamics of color naming systems, but also reveal useful properties which could be extended to artificial intelligence and robot learning. We showed that agents initialized with sparse information can come to fully understand the domain through a simple communication mechanism. Not only are they able to fully understand the domain, but they are also able to come to an understanding which is structurally similar to a system that was fully initialized. That is, the structure that is being evolved in the sparse case is not random—it is representative of some kind of empirical reality. The ability to generate systems using only sparse information would be greatly beneficial to teach autonomous robots about domains that is imperceptible to humans, such as teaching a robot to perceive and categorize light in the infrared and ultraviolet regions of the electromagnetic spectrum [6].

The implementation of this model provided insight into the dynamics of changing color naming systems with varying levels of initial sparseness. Investigating the dynamics of these gapped systems is an important step in better understanding how color naming systems evolve at very early stages of evolution.

Chapter 4

Uncovering Universal Patterns in Color Naming Using a Bayesian Nonparametric Model

4.1 Introduction

Scientific inquiry involves understanding universal properties or laws that drive natural or social phenomena. Universality can be studied using two different approaches: (i) taking micro instances as the unit of analysis to generalize to macro phenomena and (ii) starting from macro patterns to explain instances in micro phenomena. Researchers in comparative studies usually use the first approach. Yet, with recent advancements in computation power and machine learning algorithms, the second approach to universality is now more realistic. Researchers can base their machine learning models on shared universal features rather than a myriad of factors specific to the micro instances. This kind of approach reveals extant patterns without relying on prior assumptions, cultural knowledge, or predefined subgroups

and instead lets the data identify the number of "natural" groups present.

However, technical difficulty of these advanced machine learning techniques can make them difficult to apply in other fields, such as in the social sciences. In order to utilize these techniques, social science researchers would need an intimate knowledge of the social phenomenon and the algorithms, as well as the programming skills necessary for implementation. Hence, there is great need and opportunity for collaboration between computer and social scientists on the application of machine learning to social science topics. Machine learning has the potential to revolutionize social scientific inquiry and open doors to new and important discoveries, just as it had a great impact on predictive algorithms used in online commerce or media.

This work lives at the intersection of the social and computer sciences by employing advanced machine learning techniques informed by key features of color naming systems; thus, taking the second approach to universality. Color is a useful example given the universality of its concept and its well-established scholarship—spanning decades and disciplines. Past research has approached color naming by drawing inference about shared meaning across languages through the examination of meaning within languages. Conversely, we develop a methodology which characterizes the data without reference to language or culture and use unsupervised machine learning, namely the Beta-Bernoulli Dirichlet Process Mixture Model with Variational Inference, to provide an efficient means for carrying out the second approach.

4.1.1 The Study of Color Naming

The long-standing history of color naming in academia, began with Gladstone's seminal work on ancient Greek color terminology in 1858 [36]. His work was then extended by other 19th century researchers to additional ancient languages. In 1969, Berlin and Kay published

their book *Basic Color Terms: Their Universality and Evolution* [10], reigniting the study of color naming. Attempting to validate the claims made in their book, Berlin, Kay, and colleagues collected a data set known as the *World Color Survey* (WCS) [66, 23]. Basic color terms, as defined by Berlin and Kay, are the smallest set of color category names with which a person could name the entire color space. This set is determined by evaluating each color term in a language against a series of linguistic criteria (see Section 1.1). The WCS was undertaken to validate the claims made in [10] and consequently produced a famous data set which includes color naming data for many languages worldwide (see Section 1.1). This data has been analyzed with a variety of methodologies, including mathematical and machine learning methods.

4.1.2 Quantitative Approaches to the World Color Survey

At the time of their inception, the discussion and analysis of basic color terms was predominantly based in linguistics and anthropology. In recent years, though, there has been an increase in the application of quantitative methods and approaches to the WCS data. The advantage of employing such quantitative methods is the ability to perform analyses independent of cultural and linguistic features of the language.

4.1.2.1 Mathematical Approaches

Fider et al. [31] revisited the original work of Berlin and Kay [10] by redefining color term *basicness* mathematically; only taking into consideration the naming and foci data from the WCS. This mathematically-based definition was a significant advance because it provided a way to evaluate basicness independent of syntactic or semantic features of the language. They used a color strength function to classify all terms used by participants as *basic*, *potentially basic*, or *not basic*. They found that an overwhelming majority ($\sim 92\%$) of their

classifications were consistent with the classifications Kay et al. [66] made using Berlin and Kay's definition of basicness [10]. These results provided quantitative evidence in support for the ideas of Berlin and Kay which were originally formulated using linguistics.

Additionally, Fider et al. amended their category strength function to draw further inference on the WCS data set—namely, to determine the location of category boundaries [30] and to investigate differences among male and female participants [32]. Their successful application of these mathematical methods not only reinforced the original findings of the WCS, but also performed additional analyses which would have been nearly impossible to do otherwise. Also, the application of these methods provides a way to identify points for further analysis, such as color regions with low salience or cultural factors which drive the observed male-female categorization differences.

4.1.2.2 Simulation-based Approaches

Regier et al. [111] created a simulation-based model in order to formally test Jameson and D'Andrade's claim that the apparent universality of color naming could be attributed to naturally salient regions caused by irregularities in the color space [51]. Their model was centered around a measure called *well-formedness*. The goal of the simulations was to generate optimal partitions of the 330 Munsell color chip set, which could then be compared to the empirical partitions given by WCS participants. These comparisons yielded two main conclusions: (i) that most WCS participant data reflected optimal or near-optimal partitions of the color space, and (ii) that rotations of individual naming structures lowered the overall well-formedness of the system, indicating that these observed partitions are tied to specific regions of the color space. Their results were found to be consistent with Jameson and D'Andrade's hypothesis.

Agent-based simulation models have been used previously by Baronchelli, Puglisi, Loreto,

and colleagues to study the mechanisms and factors influencing the evolution of color naming conventions [103, 80, 132]. One such study compared two types of "worlds"—one representing human agents (using WCS data) and one representing neutral agents (using randomized data)—by engaging populations within the worlds in a *Category Game* [4]. They found that the worlds of human agents evolved to systems which were less dispersed, or more tightly clustered, than the neutral agents. This finding was presented as evidence for the existence of universality in color categorization.

Building on the work spearheaded by Jameson and Komarova [73, 72, 58, 59, 60, 52, 93, 102], Gooyabadi et al. modified a evolutionary game theoretic model of color category evolution [73] to be initialized with WCS seed data [39] (see Chapter 2). In this model, agents engage in a series of pairwise interactions with other agents and update their naming strategies via simple reinforcement learning based on the outcomes of these interactions. These pairwise interactions are performed sequentially until the population reaches a "stable" solution. Agents were initialized using naming data from a WCS participants and all agents from a WCS language were grouped into a population. 108 of the WCS languages were initialized as one of these simulations. Results showed that the model could successfully draw out the evolution of these color naming systems without significantly altering the system's underlying structure. The paper served as an initial demonstration of the use of evolutionary models to evaluate theories of color category evolution.

4.1.2.3 Machine Learning Approaches

The application of machine learning methods to the WCS data has been conducted primarily by Brown and Lindsey. They performed the k-means clustering algorithm on both the individual terms used [77] as well as individual participant systems [78]. The clustering of the color terms revealed a constrained set of color categories—easily equated with the English color categories and some composites—with which almost all languages could partition the

color space. The use of k-means to cluster the set of categorization systems given by individual WCS participants revealed universally occurring *motifs*. They found that (i) motifs were widespread and present in many unrelated languages, pointing to their universality, and (ii) that there was a surprising level of diversity of motifs within languages. Both of these studies revealed universal properties of the WCS data set, some of which were consistent with the claims of Berlin and Kay [10]. These studies showed the variety of ways machine learning methods could be applied to this data set and their ability to draw conclusions about the universality (and even the possible evolutionary trajectory) of color naming systems.

We follow in the same vein as Lindsey and Brown [78] in that we hope to discover possible universal structures through a clustering of participant data. However, our approach differs from previous literature in that we endeavor to perform a clustering of the data without any assumptions about the level of complexity of the model. This can be achieved through the use of nonparametric variational Bayesian inference methods to approximate the parameters of a generative, unsupervised infinite mixture model.

4.1.3 Bayesian Nonparametric Models

A limitation of the standard clustering algorithms (e.g. k-means, expected maximization, Gaussian mixture model) is that they assume a fixed number of clusters; that is, the number of resulting clusters is a parameter passed to the model. However, in many cases, the true number of clusters is unknown. One possible solution to this limitation is to execute the same algorithm, varying the number of clusters from $a \in \mathbb{Z}^+$ to $b > a$, then use an optimization technique—such as the "elbow" method or the gap statistic—to determine the optimal number of clusters. This solution, however, has constraints including the fact that these optimization methods are subjective and different methods do not always agree on the prediction for the optimal number of clusters.

Infinite mixture models, on the other hand, assume an infinite number of latent clusters which enable them to endogenously determine the number of clusters based on the data. An infinite mixture model is an instance of a Bayesian nonparametric model (BNP) applied in the context of a clustering task. The primary advantage of Bayesian nonparametric models is that the data determines the complexity of the model rather than complexity being ascertained via model selection *ex post facto* using sets of comparative models with varying levels of complexity [34, 98]. BNP uses a single, adaptive model that allows complexity to increase as more observations are introduced to the model. Consequently, BNP reveals the hidden (latent) distribution of clusters and then uses inference methods to determine the model that the data was most likely drawn from. The model we employ is the Beta-Bernoulli Dirichlet Process Mixture Model with Variational Inference (BBDP) [96, 48, 17].

4.1.3.1 Beta-Bernoulli Observation and Mixture Models

Suppose we have data set $X = \{X_1, ..., X_N\}$, where each observation X_i is a binary vector with D dimensions representing D attributes of an observation. An entry $x_{id} = 1$ if X_i has the attribute d and $x_{id} = 0$ otherwise. If we let θ be the mean of the Bernoulli distribution, then the Bernoulli likelihood can be written generally as:

$$P(x|\theta) = \theta^x(1-\theta)^{1-x} \tag{4.1}$$

Using this form, the probability density for each observation X_i can then be computed by:

$$P(X_i|\theta) = \prod_{d=1}^{D} \theta_d^{x_{id}}(1-\theta_d)^{1-x_{id}} \tag{4.2}$$

where θ is a D-dimensional vector with entries θ_d for $d \in \{1, ..., D\}$ represent the probability that an observation has the attribute d.

The conjugate prior to the Bernoulli distribution is the Beta distribution with parameters β_1 and β_2. Therefore, the prior, $P(\theta)$ can be given by the following function:

$$P(\theta) = \frac{1}{\mathbf{B}(\beta_1, \beta_2)} \theta^{\beta_1 - 1} (1 - \theta)^{\beta_2 - 1} \tag{4.3}$$

where the Beta function $\mathbf{B}(\beta_1, \beta_2)$ serves a normalization constant and β_1, β_2 are shape parameters that determined based on prior beliefs or existing knowledge. We only search the portion of the parameter space where $\beta_1, \beta_2 \in (0, 1)$ because the shape of the beta distribution is biased towards the bounds of its domain, 0 and 1, when $\beta_1, \beta_2 < 1$. This behavior is useful when drawing priors for the a Bernoulli mixture model.

A Beta-Bernoulli mixture model can be defined by a mixture of K Beta-Bernoulli distributions. In order to identify which of the K distributions each data point X_i was drawn from, we introduce a latent variable $Z = \{Z_1, ..., Z_N\}$. For each $Z_i \in Z$, Z_i is a K-dimensional vector which has exactly one entry equal to 1, corresponding to the cluster assignment of X_i. Each of the K distributions in the mixture model has a corresponding weight, represented by $\pi = \{\pi_1, ..., \pi_K\}$, such that $\sum_{k=1}^{K} \pi_k = 1$. Therefore, the distribution of the latent variable Z conditioned upon its weights π is:

$$P(Z|\pi) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}} \tag{4.4}$$

and the Bernoulli likelihood can then be formalized as:

$$P(X|Z, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} P(X_i|\theta_k)^{z_{ik}} \tag{4.5}$$

Figure 4.1 depicts a graphical model representation of the Beta-Bernoulli mixture model.

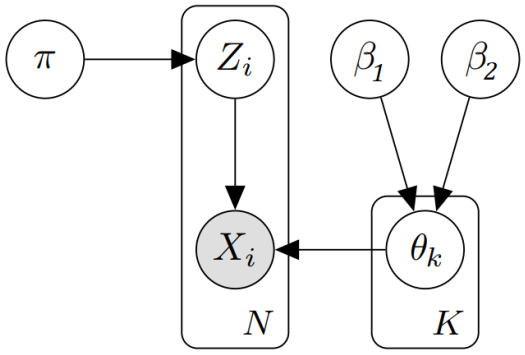

Figure 4.1: A graphical model of the Beta-Bernoulli mixture model.

4.1.3.2 Dirichlet Process Mixture Model

The Dirichlet Process (DP) is a nonparametric prior for infinite, discrete distributions. Therefore, the DP mixture model is able to cluster exchangeable data points without determining the number of clusters *a priori* by assuming an infinite number of latent clusters. For this reason, DP mixture models are synonymously known as infinite mixture models. These processes are commonly used in Bayesian nonparametric methods because they allow the number of clusters to grow as more data points are introduced to the model.

A DP can be thought of as a distribution over distributions. Suppose G is a Dirichlet process, then

$$G \sim DP(\alpha, G_0)$$

where $\alpha \in \mathbb{R}^+$ is called the dispersion parameter and G_0 is the base probability distribution. Draws from the process G are taken according to the following algorithm:

1. Assume there are $X_1, ..., X_N$ observations and k unique values for the variable K (which represent k clusters present at the time).

2. For observation X_i, with probability $\frac{\alpha}{N-1+\alpha}$, a new draw is taken from G_0 (i.e. X_i is assigned to a new cluster).

3. With probability $\frac{n_k}{N-1+\alpha}$, where n_k is the number of observations currently in cluster

k, X_i joins cluster k.

4. Each observation is iteratively assigned to a cluster until all N observations have been grouped. Cluster assignments are stored in the latent variable Z where $Z_i \in Z$ is a K-dimensional vector with the k-th element being equal to 1 (corresponding to datum X_i being assigned to the cluster k) and all other elements equalling 0.

The end result of this process is then a distribution over the partitions of the data X, which serves as a prior over the class assignment vector Z. Some common analogies used to describe the DP are the Chinese Restaurant Process, the Stick-Breaking Construction, and a modified version of Polya's Urn Scheme. A graphical model for the DP mixture model is presented in Figure 4.2.

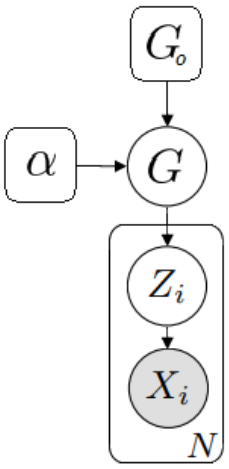

Figure 4.2: A graphical model of the Dirichlet Process mixture model.

4.1.3.3 Variational Inference

Due to the complexity of the statistical models in Bayesian nonparametric methods, many of the resulting integrals become intractable and thus require other techniques to approximate the parameters of the model. One such family of techniques is called *variational Bayesian inference*. Variational inference can be used as a way to (i) estimate the model's posterior

distribution or (ii) to compute an evidence of the lower bound (ELBO), which is then used for model selection. The intuition behind (ii) is that the higher the computed marginal likelihood of a model is, the higher the probability that the data was generated by that model. Therefore, the model with the highest ELBO is selected as the most appropriate model, given the data. In this paper, we use variational inference for the purpose of computing the lower bound of the marginal likelihood.

Given a set of unobserved variables Z and a data set X, the posterior distribution can be approximated by the variational distribution Q:

$$P(Z|X) \approx Q(Z)$$

The aim of variational inference is to minimize the distance between the true posterior $P(Z|X)$ and the approximated distribution $Q(Z)$ and thus seeks to find the Q which minimizes this distance. The distance between distributions P and Q is most often formalized using Kullback-Leibler Divergence (KL-divergence), defined as

$$KL(Q \parallel P) = \sum_{Z} Q(Z) \log \frac{Q(Z)}{P(Z|X)} \tag{4.6}$$

This function can be rewritten and rearranged to yield

$$\log P(X) = KL(Q \parallel P) - \mathbb{E}_Z \left[\log Q(Z) - \log P(Z, X)\right] \tag{4.7}$$

$$= KL(Q \parallel P) + \mathcal{L}(Q) \tag{4.8}$$

The term $\mathcal{L}(Q)$ is called the Evidence Lower Bound (ELBO). Maximizing $\mathcal{L}(Q)$ will reveal the Q which minimizes the KL-divergence since $\log P(X)$ is fixed with respect to Q.

4.2 Methods

Color naming is an inherently linguistic task as it involves assigning a lexical term (category name) to a stimulus (color). To discover universal patterns in color naming using only universal features, it is prudent to abstract away from each language's color terms. The universal features influencing color naming include language, biology, and the topology of the stimulus space. Therefore, our model is tailored to accommodate these universal features.

The color domain is a continuous space on which the human eye can discriminate millions of colors. Communicating about color requires a language to bestow its speakers with a set of color terms which categorizes the perceived color space. Language allows us to partition the continuous color space into a manageable set of discrete categories (i.e. basic color terms [10]) [78, 117, 103]. Due to the competing nature of the categories at the boundaries, name assignments near these regions become less certain. Therefore, much of the individual variation in color naming systems occurs at the category boundaries. We transform the data into a form that captures these differences in individuals' category boundaries in order to incorporate as much of the structural nuance into the model as possible.

The WCS *naming task* (see Section 1.1) was a forced naming task, meaning that participants were required to name every color chip regardless of the chip's salience. Regions of low salience have been exhibited in multiple studies [79, 141, 39]. Consequently, especially when assigning a name to a color chip in these regions, participants rely on their own internal perceptual processing, rather than linguistic conventions, to make a judgment. These visual processes are governed by biological features which are universal across most people. Since these individuals are drawing from processes which are shared amongst most human beings, we can treat the participants as equal and anonymize them from their native language. To do so, we use features of the color space to transform the data into a form that is agnostic of their culture and language.

The stimulus space used in the color naming task is a discrete set of color chips sampled from a Mercator projection of a 3D color solid with its own unique geometry [51]. In this color space, colors which are closer in proximity will be perceived more similarly. This type of stimulus space is fundamentally different than one where each stimulus is independent from other stimuli (e.g. pictures of dogs) or one that is directly related to each other (e.g. tracking movement of a ball in a video clip). There is both an underlying relationship between color chips in the space yet each chip is a unique stimulus. This distance-dependent relationship between adjacent color chips is used as a rationale for transforming each participant's color naming data into pairwise judgments (see Section 4.2.2). This transformation abstracts away from their linguistic origin, enabling a language-independent way of drawing on similarities between participants to uncover universal patterns.

4.2.1 Data

This study uses *naming task* data (see Appendix 1.1) gathered from participants of the WCS [23]. Of the 110 languages surveyed, four languages[1] are omitted from our data set due to data collection and transcription issues. Additionally, the 10 achromatic chips (column A0–J0 in Figure 1.2) are omitted from our data set due to their disjointed nature from the 320 chromatic chips. Therefore, "the data" will henceforth refer to the collection of *naming task* data for 106 languages (2,552 individuals) on the 320 chromatic chips in the color grid.

4.2.2 Transforming WCS Participant Data

The algorithm is designed to be agnostic of the languages the participants originally came from. Participants are dissociated from the specific color terms they used by transforming

[1]The four languages omitted are: Huastec (Language 45), Mampruli (Language 62), Tarahumara-C (Language 92), and Tarahumara-W (Language 93).

the naming task data into an n-dimensional binary features vector. The vectors are constructed by comparing the names assigned to neighboring color chips. Chips in the chromatic component of the WCS color grid (B1–I40 in Figure 1.2) are chosen one at a time to be the *reference chip*. The name for the *reference chip* is compared to the name of one of its 4 vertically and horizontally adjacent neighbors (or 3 in the case of chips located on rows B and I[2]). Each element of the binary vector represents one of these pairs. An index in the vector is given a value of 1 if the two color chips being compared have the same name and 0 if they have different names (see Figure 4.3).

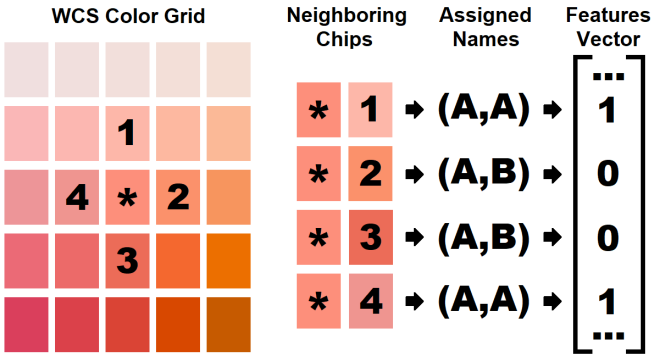

Figure 4.3: An illustrated example of the process to transform a WCS naming system into binary features vector. First, a color chip is picked to be the *reference chip* (represented by *). Second, the *reference chip* is compared to all of its adjacent neighbors. If the two neighboring chips were given the same name by the WCS participant, then the element in the features vector is given a value of 1, otherwise it is set to 0.

Transforming the *naming task* data yields a set of 2,552 data points each with 2,320 binary attributes. In order to prevent the model from needing to parse through an excessive amount of data and since this pairwise judgment is reflexive, redundant pairs are omitted from the features vector; that is, (chip i, chip j) is included in the vector, but not (j, i).

[2]Color grid is a Mercator projection of the Munsell color space, so the ends of the rows are considered connected but the tops and the bottoms represent the poles of the 3-dimensional solid. Therefore, chips in rows B and I do not have neighbors in the north and south direction, respectively.

4.2.3 Model Implementation

We use the Python package BayesPy [84] to implement variational inference methods. It has built-in functions for performing variational Bayesian inference on conjugate exponential family models. BayesPy approximates infinite dimensional distributions, such as the Dirichlet process, by setting the maximum number of clusters K to a value much higher than the number of expected clusters. We initialize $K = 100$ clusters and consistently find the number of resulting clusters $K^* < 100$, indicating that the results are driven by the data and not the upper bound.

4.2.3.1 Selecting Hyperparameters

Though one of the main benefits of implementing a model using Bayesian nonparametric methods is the ability of the model to freely determine parameters' values throughout the training process, these models still contain variables which need to be exogenously determined. These variables are called *hyperparameters*. Several methods are commonly used in order to search the parameter space for set of hyperparameters which will yield the most "optimal" result, such as grid search, random search, Bayesian optimization, and evolutionary optimization [8, 7]. We chose to use random search to estimate values for our model's hyperparameters.

A naïve search method is employed instead of one which actively searches for an optimum (e.g. Bayesian or evolutionary optimization) because the more simplistic approach was found to be sufficient for the purposes of this study. Hence, random search is used to search the parameter space for an estimate of the optimum. The optimal parameters are the determined by finding the combination which yields the highest ELBO.

Several sets of 100 random initializations were run at a time in an effort to determine general

regions of optimality. This revealed a broad pattern. Higher values of β generated higher ELBOs whereas α did not appear to have much of an influence on the value of the ELBO (see Figure 4.4). This finding is consistent with the fact that the model is more sensitive to the beta distribution hyperparameters than the Dirichlet process concentration parameter [96].

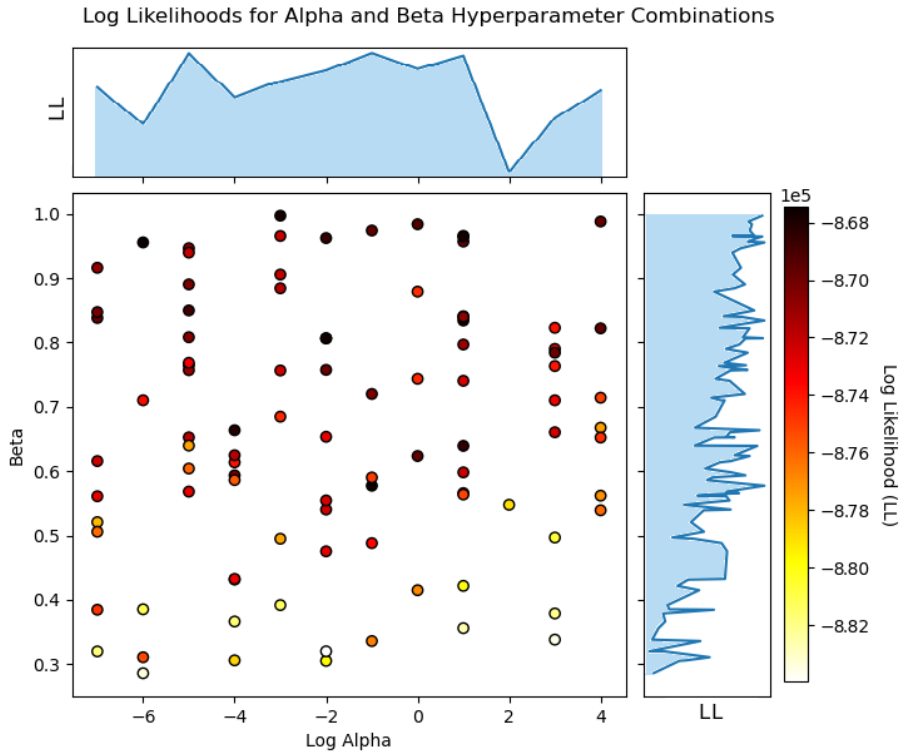

Figure 4.4: Scatter plot representing 100 random initializations of the BBDP. The x-axis represents the α parameter (Dirichlet process concentration parameter) in log units. The y-axis represents the β hyperparameter (β_1, β_2 of the beta distribution). The darker the color of the data point, the higher the ELBO of the algorithm run using those hyperparameters.

Therefore, based on the pattern obtained from running multiple set of initializations and precedence from previous literature, the hyperparameters selected for the model were $\alpha = 1000$ and $\beta = \beta_1, \beta_2 = 0.9$. These values were used to instantiate the model which generated the results reported in Section 4.3.

4.2.4 Cluster Analysis

Specific and novel methodologies are developed to perform within and between cluster comparisons. These methods are useful for visualizing the resulting clusters and conducting customized statistical and quantitative analysis.

4.2.4.1 Centroids

The *centroid* is designed to provide a singular representation of the group. Centroids are constructed by taking the modal term used by the group for each color chip. By constructing the centroid in this manner, it becomes the representation which minimizes distance from every other member of the group.

Given that a cluster can consist of participants from more than one language group it is necessary to first create a translator that maps all color terms into one common "language". This translator is constructed by performing k-means clustering over the terms used by all participants in the group. Terms that get clustered together are then considered equivalent (i.e. refer to similar regions of the color space). This is the method established in [78] with the exception that the gap statistic is exchanged for the average silhouette method [116, 64] to determine the optimal number of clusters. First, we generate one 320-dimensional binary vector for each color term used by every WCS participant in the group. The i-th element of a vector is given a value of 1 if color chip i is named using that term. Second, the distance between two color terms, x and y, is computed using Pearson's R, $d(x, y) = 1 - \rho(x, y)$. We then use the following procedure to determine how many terms should be included in the universal translator:

1. Perform k-means clustering for values of k ranging from 2 to the minimum number of terms used by any participant in the group.

2. For each k, calculate the average silhouette over all observations (i.e. terms used by participants).

3. The k value which yields the maximum average silhouette is considered the appropriate number of clusters.

4. Color terms that are assigned to the same cluster are inferred to by synonymous with each other and are thus mapped to a singular, common term.

Once all participants in the group have been translated into the common "language", the modal map can be constructed by taking the modal term for each color chip, resulting in the centroid representation of the group.

4.2.4.2 Boundary Heatmaps

To reveal the underlying strength of the category partitions, a more informative representation of the resulting clusters is depicted through *boundary heatmaps*. While centroids reveal the color space partition, boundary heatmaps show the strength or level of agreement over the partition. This additional information is imperative because it discriminates between two similar modal maps by revealing the regions within the partition that are most salient to the cluster.

The boundary heatmap assigns a value to each color chip in Figure 1.2 by computing its *boundary probability* (i.e. the likelihood that a color chip is located on the boundary of a color category; see Equation 2.10 in Section 2.3.2.1). This represents the strength of the category boundary at that chip.

4.2.4.3 Schematic Similarity

Schematic similarity (SS) is a tool that can be used to compare two participants' color naming data without making any assumptions about the language or culture the participants come from. This analysis is based on the partition itself and, therefore, is not dependent on the names assigned to regions of the color space. SS performs term comparisons at the participant level in an effort to preserve maximal information.

SS calculates similarity between two participants by determining the amount of overlap among their corresponding terms. Terms that refer to similar regions of the color space are mapped to each other. Once all the terms are mapped, the metric does the following to compute the amount of overlap:

1. Set the participant with fewer color terms as the *Reference Participant* (RP) and the participant with the more terms as the *Other Participant* (OP). Let n be the number of color terms used by the RP.

2. Find the best term equivalence for Term 1 used by RP using *Term Similarity*.

$$TS(T_i^{RP}, T_j^{OP}) = \frac{|T_i^{RP} \cap T_j^{OP}|}{|T_i^{RP} \cup T_j^{OP}|} \tag{4.9}$$

   where T_i^{RP} represents the set of color chips that RP named with term i and T_j^{OP} represents the set of color chips that OP named with term j.

   Let T^{OP} be the set of terms used by OP. The term used by OP which best matches T_1^{RP} (Term 1 used by the Reference Participant) is $\arg\max_{j \in T^{OP}} TS(T_1^{RP}, T_j^{OP})$.

3. Let M be a set, where each element is a tuple representing the mapped terms between RP and OP. Suppose from Step 2 that $\arg\max = k$, then the tuple $(1, k)$ would be added to M.

4. Repeat Steps 2 and 3 for Terms 2, ..., n used by RP (see Figure 4.5a).

5. Once all of RP's terms have a best match, calculate *Schematic Similarity* between RP and OP.

$$SS(RP, OP) = \sum_{(i,j)\in M} \frac{1}{n} \frac{|T_i^{RP} \cap T_j^{OP}|}{|T_i^{RP} \cup T_j^{OP}| + e_{T_i^{RP}}} \tag{4.10}$$

where $M = \{(i,j),(a,b),...\}$ represents the set of corresponding/mapped terms from RP to OP and $e_{T_i^{RP}}$ represents the error in term i.

Term error is calculated using the following formula,

$$e_{T_i^{RP}} = \sum_{k\in U} |T_i^{RP} \cap T_k^{OP}| \tag{4.11}$$

where U is the set of terms used by OP that did not get mapped to an equivalent term from RP (see Figure 4.5b).

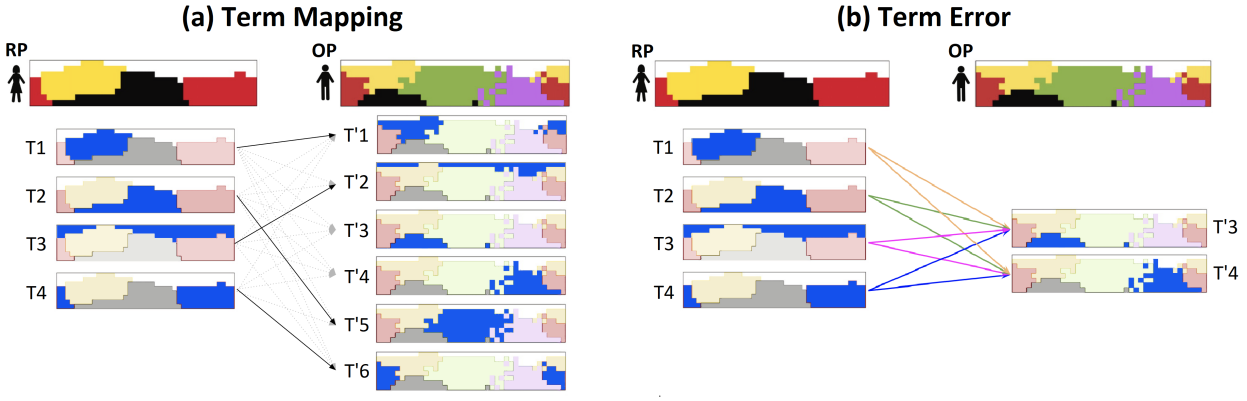

Figure 4.5: Example of (a) term mappings and (b) term error where each term is represented by the blue region in front of T for the RP or T' for the OP. In (a), left participant is RP (n=4) and the right participant is OP. Black arrows represent the term mapping (e.g $T2$ mapped to $T'5$). In (b), two of OP's terms, $T'3$ and $T'4$, are not mapped to any of RP's terms and contribute to the Term Error incurred by each of RP's terms. For instance, the error of $T1 = |T1 \cap T'3| + |T1 \cap T'4|$.

After determining the SS between two WCS participants, the distance between the participants is defined to be:

$$d(RP, OP) = 1 - SS(RP, OP) \tag{4.12}$$

This distance metric provides a mathematical quantity with which to determine the dispersion within a cluster and to calculate the distance between cluster centroids. (See Appendix B for the proof that $d = 1 - SS$ is a metric.)

4.2.4.4 Group Error

Due to the forced nature of the WCS *naming task*, regions of the color space with low cultural salience likely caused some of the observed variation in the naming systems of participants from the same language. To study within language variation, we develop *group error*. Group error is a novel measure that quantifies the diversity within each language group for post cluster analysis. With this measure, we can investigate the distribution of universal patterns across groups. Group error is a function of the number of clusters a language group is split up into and the distance between those clusters. To calculate the error for that language, participants are considered in a pairwise fashion. The distance between two participants in the same cluster is 0 while the distance between two participants in different clusters is the distance between their respective clusters. Group error is formally defined as:

$$e_g = \frac{\sum_{(i,j)} d_{ij}}{\binom{N_g}{2}} \tag{4.13}$$

where N_g is the number of participants in the group g, $i, j \in \{1, ..., N_g\}$, and

$$d_{ij} = \begin{cases} 0 & \text{if } \ell_i = \ell_j \\ 1 - SS(C_i, C_j) & \text{if } \ell_i \neq \ell_j \end{cases} \tag{4.14}$$

where ℓ_i represents label of the cluster containing participant i, ℓ_j represents label of the cluster containing participant j, and C_i and C_j are the centroids of the i and j's clusters, respectively[3].

4.3 Results

4.3.1 Model Efficacy

To determine the effectiveness of the BBDP model for the tasks outlined, it was first tested using "synthetic" data generated by the simulations described in Chapter 2 [39]. The data consisted of agent populations (representing WCS languages) with nearly identical naming systems; therefore, we expected the BBDP to place members of these populations in the same cluster. We found that 84% of languages had a group error of 0, meaning that all members of the language were assigned to the same cluster. This demonstrates that the model successfully identified similarities in naming structure and accurately grouped them together.

After confirming that the BBDP successfully clustered the "synthetic" data, the task was then performed using our data. The resulting clusters were compared against three baseline groups: (i) a random grouping of participants; (ii) an existing grouping of participants (e.g. by the number of basic color terms (BCTs) identified for the languages in the WCS analysis [69]); (iii) a representative subset of participants. We generated SS histograms to compare the level of similarity within and between the aforementioned groups (see Figure 4.6).

For (i), we randomly assigned participants into 10 groups[4] to represent "psuedo-BCT stages".

[3]Since group error is a novel contribution, we used an established diversity measure—Simpson's Index—to validate it. We found Simpson's Index and group error to have a highly significant correlation of $\rho = 0.9$ ($p\text{-}value \approx 0$). Thus, we have evidence that group error is capturing the intended phenomenon.

[4]WCS systems identified to have anywhere from 3 to 12 BCTs; thus, 10 random groups to represent 10

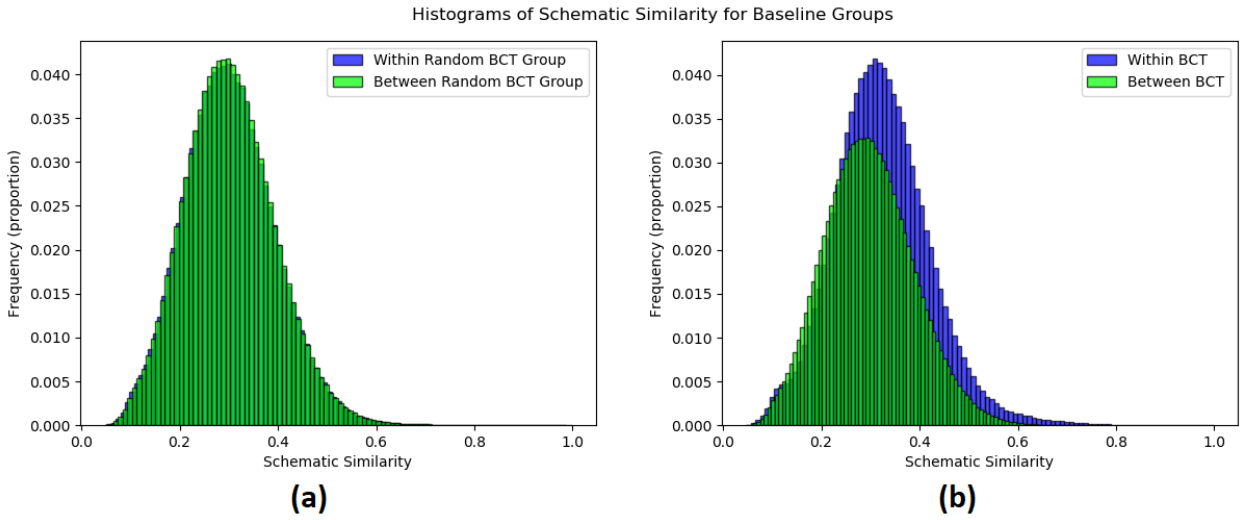

Figure 4.6: Distributions representing within group SS and between group SS for (a) randomly assigned pseudo-BCT groups and (b) WCS identified BCT stages.

The within group histogram had $\mu_w = 0.30$ and $\sigma_w = 0.092$, and the between group histogram had $\mu_b = 0.297$ and $\sigma_b = 0.092$ (see Figure 4.6a)—the two distributions were almost perfectly concurrent. This pattern suggests that the random groups were not capturing any inherent structure among participants.

For (ii), participants were assigned to groups according to the number of BCTs identified in their language (e.g. all members from languages with 3 BCTs were grouped together). This grouping is based on properties used in previous studies [10, 66] to draw comparison between naming systems. The within BCT group histogram had $\mu_w = 0.324$ and $\sigma_w = 0.098$, and the between BCT group histogram had $\mu_b = 0.297$ and $\sigma_b = 0.0899$ (see Figure 4.6b). This slight positive shift in the within group SS away from the between group SS indicates that the number of basic color terms captured some level of the similarity among the color categorizations.

Comparing the results of the first two baselines groups to the clustering results generated by our model helped to evaluate how well the inferred clusters captured the structural similarity

BCT groups.

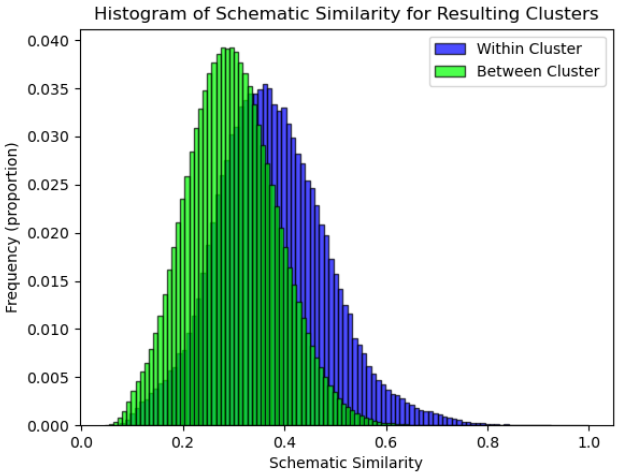

Figure 4.7: Distributions representing within group SS and between group SS for resulting clusters from the BBDP.

between participants. Our clusters were found to be better representations of both baseline group with a higher within group SS (see Figure 4.7). The within cluster distribution had a $\mu_w = 0.384$ and $\sigma_w = 0.111$, where as the between cluster distribution had $\mu_b = 0.298$ and $\sigma_b = 0.089$. The more dramatic positive shift in the within cluster distribution away from the between cluster distribution indicates that the model is capturing similarities and nuances in the structure of these color naming systems that was not being captured by either of the baseline groups. Comparing results from clusters to the third baseline group ($\mu_w = 0.404$, $\sigma_w = 0.130$; $\mu_b = 0.302$, $\sigma_b = 0.0919$) showed little difference, further validating the BBDP's performance.

4.3.2 Clustering Results

One run of the model using parameters $K = 100$, $\alpha = 1000$, and $\beta_1, \beta_2 = 0.9$ yielded a total number of 88 inferred clusters. Of these 88 clusters, 85% of participants were placed in 18 clusters, visualized by centroids and boundary heatmaps in Figure 4.8. The remaining clusters each contained less than 1% of the data set and were removed from subsequent

cluster analysis.

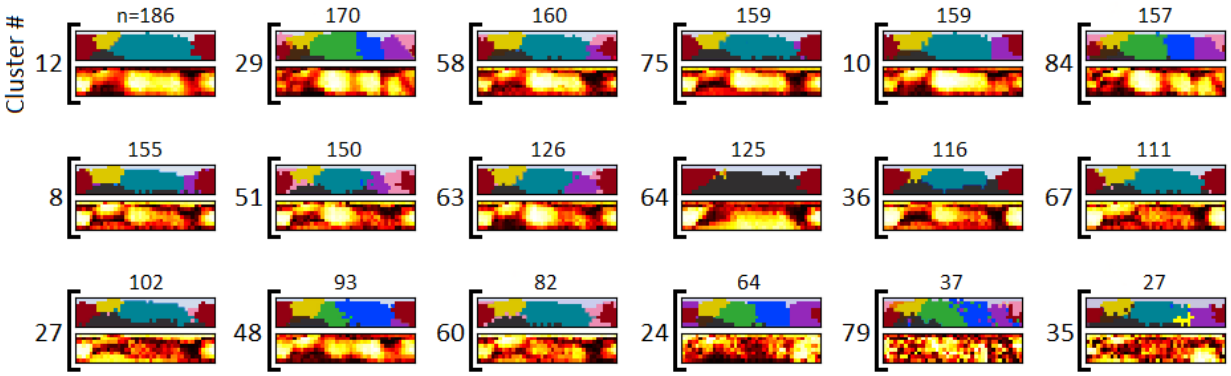

Figure 4.8: Cluster representations for the 18 analyzed clusters with number of members (n=) shown on the top and cluster number (#) on the left-hand side of each group for reference in Figure 4.10. Clusters are visualized through their centroid (top) and boundary heatmap (bottom). Colors in the centroid are used only to distinguish lexical terms and loosely denote the hue of the underlying region—they are false colors. In the boundary heatmaps, the darker the color, the higher the likelihood that the corresponding chip is located on a category boundary.

The average SS across all cluster means was 0.342 with a the standard deviation was 0.067. In comparison, the average SS for clusters versus WCS language groups were comparable, but the average language standard deviation ($\sigma = 0.080$) was higher than for the clusters. The smaller standard deviation of resulting clusters demonstrates that the resulting clusters from the BBDP were more tightly clustered than the language groups. This provides further evidence that the clusters are better representations and are better capturing the similarity in structure of the naming systems.

Examining each cluster, we found similarities across languages (i.e. multiple languages were represented in a cluster) as well as diversity within languages (i.e. participants from the same language did not necessarily get assigned to the same cluster). The average group error across all languages was $\mu_{e_g} = 0.314$ and was found to be statistically significant ($t = 35.40$, $p\text{-}value \approx 0$), meaning languages typically could not be characterized by a single naming pattern. This is corroborated by the fact that participants from a language were found in 6.83 different clusters on average (median=7, mode=7, 8). The range of this value,

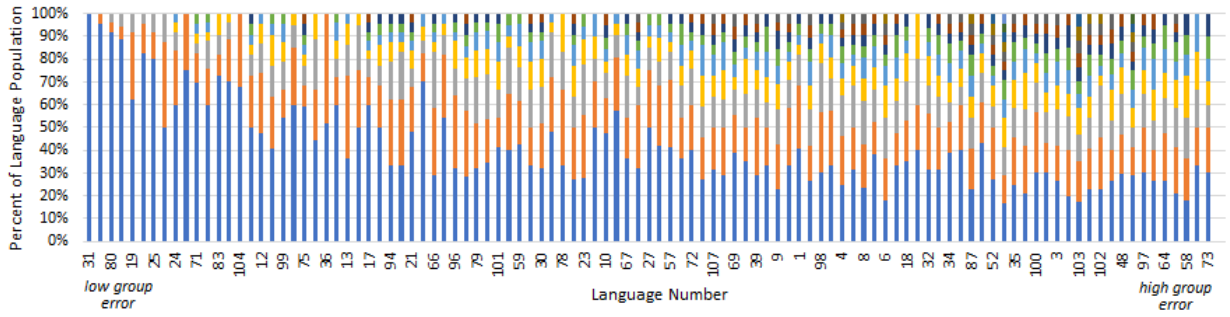

Figure 4.9: Distribution of each language across the inferred clusters ordered along the x-axis from lowest to highest group error. Each bar is a language group, each colored segment represents the clusters the members were placed in, and the size of the segment is proportional to the number of participants in the corresponding cluster. For example, all members of Language 31 (far left) belong to the same cluster while Language 73 (far right) has members dispersed amongst seven clusters.

however, varies from 1 to 13 (see Figure 4.9), highlighting that languages have varying levels of diversity. Group error and the number of clusters that the languages get divided into are positively correlated ($\rho = 0.712$, $p\text{-}value \approx 0$; see Figure 4.9). Observing the distribution of the languages across the resulting clusters provides insight into the structures (and possible evolutionary processes) present in the language at the time of collection. For example, some languages, may be characterized by a single, prominent color naming system (e.g. Language 74 [Múra Pirahã] has all but one member belonging to Cluster 36, see Figure 4.10a) or by a few strong systems (e.g. Language 85 [Seri] has members dispersed between three main clusters [12, 10, and 29] with two in other clusters). The low group error of Language 74 points to high levels of similarity and cohesion, implying that conventions associated with its lexicon are well-established within the population. This could be an indicator that Language 74 has reached a point of stability within its evolutionary stage. Conversely, the diversity present in Language 85 points to poorer lexical consensus among its speakers. This diversity could be influenced by different factors: (i) gender differences [32], (ii) weak notions of color [76, 79], or (iii) a transition from one evolutionary stage to another [78] (Cluster 12 contains the individuals without a "pink" term whereas participants in Clusters 10 and 29 possess a term for "pink"—this could point to the introduction of a new term into the system). While

the methods described here are certainly useful in examining broad trends and identifying points for further analysis, the linguistic and anthropologic tools to draw such conclusions are beyond the scope of this paper.

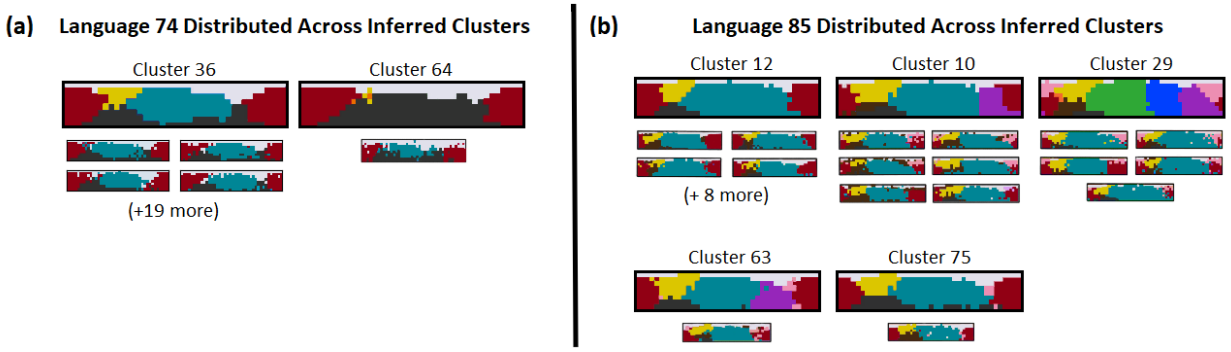

Figure 4.10: Distribution of language participants across clusters from a language with (a) low group error and (b) high group error. Larger colored rectangles represent the centroids of the cluster. The smaller colored rectangles below the large ones are the name assignments for each language participant found in that cluster. As in Figure 4.8, the colors depicting the categories are false colors.

4.4 Discussion

The primary aim of this research is to provide an initial demonstration of Bayesian nonparametric methods being applied to color naming data. Building on previous work [77, 78] which used parametric clustering algorithms, we investigate underlying features of color categorization systems with a particular nonparametric method, BBDP. This novel approach results in findings which are in agreement with the existing literature.

The largest commonality between the results presented here and previous conclusions is that the 110 WCS languages can be classified into a small number of patterns. This fact coupled with the widespread nature of these clusters across unrelated languages suggests the existence of universal processes which govern color naming systems. The inferred clusters from the BBDP resemble motifs previously identified by [78]—namely, they are representations of

shared patterns across individual naming systems. Another finding consistent with past research is the significant amount of diversity that exists even among individuals from the same language [78, 137, 32, 39]. Group error finds that most languages have multiple patterns present ($\sim$7), highlighting that diversity cannot be predicted by language membership alone.

This work extends the existing literature by further abstracting away from both linguistic and algorithmic assumptions. As mentioned previously, one of the main draws of Bayesian nonparametric models is the ability to let the data determine the complexity of the model. The most prominent work to date which uses ML algorithms on the WCS [78] identified 3–6 motifs but pointed to the need for a different approach to capture "minority motifs" that are rare but resemble evolutionary stages proposed by Berlin and Kay [10, 65, 69]. Using BBDP yields 18 prominent clusters with some having near equivalence to the motifs, with the remaining clusters possibly being these "minority motifs". Knowing the distance between the resulting clusters in conjunction with the boundary heatmaps presents a more nuanced view of the data.

A notable contribution made by this paper is the inclusion of distance when analyzing the spread of the resulting clusters. While previous literature has looked at the distribution of "motifs" across the WCS language groups, there has not been any analysis to see how far these resulting clusters are from each other. By analyzing not only the distribution of clusters but also their proximity to one another, we can uncover more information about the structure of each WCS language's population naming strategy.

The successful application of the BBDP using the transformed binary data (see Section 4.2.2) shows that "color neighborhoods" are a meaningful criterion by which to evaluate structural similarities between individuals. Results demonstrate that using local boundary information can replicate and build on past findings. Roberson et al. asserted that the determining feature of color categories were boundaries which varied across languages [114], not universal focal colors as others had claimed [70, 69, 43, 44, 108, 109]. We similarly find

that category boundaries are a deterministic feature of these systems, but unlike Roberson et al. we suggest that these boundaries are influenced by universal factors perhaps in addition to more language-specific phenomena. Given the recent developments in tools for detecting color category boundaries [30, 39], future research can further explore the properties of these boundaries.

4.5 Conclusion

This study successfully demonstrates the application of advance machine learning techniques (e.g. generative Bayesian nonparametrics) to study social phenomena based on universal features. While these models have been widely applied to settings where number of natural categories is unknown but of great interest, such as text and image processing, its application in the social sciences has been sparse. Using the second approach to universality, we base the ML model for color naming on universal features of linguistics, biology, and the color space topology to uncover extant patterns in WCS data without relying on cultural knowledge of the language groups. Specific customization for this goal included transforming the data into pairwise-neighbor judgments based on the topology of the stimulus space and developing appropriate mathematical techniques to study results. We show that BBDP is an appropriate model to uncover universal naming patterns and find our results to be in agreement with past work demonstrating that there exist universal tendencies among color naming systems worldwide.

The methods described in this chapter could be useful to those interested in studying universality in other social domains. Additionally, those interested in further investigating the WCS can use the 18 universal patterns in this paper as a starting point. Though this chapter provides initial insights into the structure of naming patterns found across the WCS languages, it does not evaluate *why* these structures might be present. Therefore, researchers

with the appropriate cultural knowledge and tools can delve deeper in understanding the societal and cultural processes present in each language. This further exploration could answer questions such as: Does a person's role in society (e.g. hunter versus gatherer) affect their color categorization? Do men and women have different vocabularies? Do languages in proximal geographic regions share structural similarities? Does this apply also to languages with the same temperate environment?

Another avenue for future research is to compare our results against other Bayesian nonparametric models. The model described here serves as a preliminary approach which, given its success, provides an avenue to test the data using more complex algorithms. Testing these complex models could be useful in uncovering other facets of color categorizations not revealed by this particular approach. Additionally, using such algorithms would result in fewer small, outlying clusters which would diminish the number of data points removed from analysis. Though there are points of expansion, the results presented here are a satisfactory introduction of nonparametric models to color naming.

Chapter 5

Determining Individual Variations in Color Perception Through Flicker Sensitivity

5.1 Introduction

The methods and results presented in this chapter are from a pilot study which was designed to investigate variations in color perception among individuals with different color vision phenotypes using a flicker sensitivity task. Results from a full-scale experiment are forthcoming. The conclusions drawn from this pilot study, however, provide valuable intuitions and predictions regarding the potential relationship between flicker sensitivity and an observer's color vision phenotype. Insights drawn from this relationship will help reveal the effects that artificial illuminants (e.g. LEDs) are having on observers' color viewing experience and will inform the production of future illuminants which are both energy efficient and faithful in color rendering ability.

5.1.1 Photopigment Opsin Genotypes

Just as many physical features vary from person to person, so does the genetic make-up of human color vision. Therefore, the colors that we assign to a certain scene perceived in the real world are not universally homogenous among all people but are, rather, unique color experiences that are governed by our genetics and viewing circumstances [62]. The genetic sequences responsible for our subjective viewing experience are referred to as photopigment opsin genes. These genes determine how human retinas respond to the visible light range of the electromagnetic spectrum.

Typically, human beings have genes that allow the expression of three retinal photopigment classes. It is presumed that all of these *trichromats*—or, individuals with three cone classes—have a similar organization of color vision mechanisms (though there does exist normal photopigment variations which may cause observers to have different color appearance spaces). However, there are individuals who have atypical genetic sequences or events that give rise to anomalous genes, gene copies, or deletions which result in observers who express either one fewer or one additional cone class compared to the normal trichromat. Individuals who express only two retinal photopigment classes are referred to as *dichromats*. Individuals who potentially express four retinal photopigment classes are referred to as *potential tetrachromats* [55, 62] (see Figure 5.1).

Even among the more common trichromacy, there are variations in the absorption spectra of a trichromat's three cone classes [115]. Identification and specification of genetic bases underlying retinal photopigments is fairly recent development. For example, Nathans et al. discovered an anomaly in the red cone [95]. They concluded two different phenotypes could be caused by the presence of a hybrid gene in an observer's red cone crossing-over with normal green pigment genes. This anomalous red curve was found to be between the normal red and green curves. In the extreme anomalous case, the red curve had shifted far to the

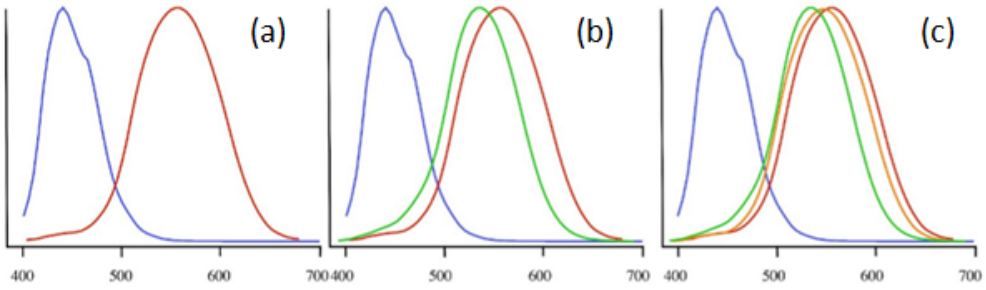

Figure 5.1: Example cone response curves of (a) a dichromat (specifically, a deuteranope—missing M-cones), (b) a trichromat, and (c) a potential tetrachromat. The x-axis represents the visible light range of the electromagnetic spectrum (400–700 nm) and the y-axis represents the response of each cone class for any given wavelength of light.

left such that the interval between the anomalous red and normal green curves was very small [95]. A consequence of this shift in the red curve is a very distinct difference in color matching between different observers.

"Dichromacy and anomalous trichromacy are caused by unequal homologous recombination within the tandem array of genes that encode the red and green visual pigments" [90]. Therefore, explanations for this anomalous shift of the red curve can be articulated by looking at color vision genotype. "The largest shifts in λ_{max} are produced by substituting alanine for threonine at codon 285 (ca. -14 nm), phenylalanine for tyrosine at position 277 (ca. -7 nm), and alanine for serine at position 180 (ca. -4 nm) (Merbs and Nathans, 1992a)" [119]. Variations among amino acids at these codons are present among both M-cone and L-cone photopigment genes, though variations in the L-cone are much more prominent. For this study, only alanine and serine variations at position 180 were taken into consideration.

5.1.2 Potential Human Tetrachromacy

Among the many photopigment opsin phenotypes exists anomalous tetrachromacy. Tetrachromacy is the condition of an organism to possess four cone cells in their eyes and is more commonly found in species such as birds, fish, amphibians, reptiles, and insects. It has

been hypothesized, though, that some human individuals have the potentiality to express this extra cone class. These individuals, who are most likely women, have the potential to express an additional photosensitive pigment class to the usual three, and are thus referred to as *potential tetrachromats* [62]. The existence of a fourth photoreceptor class in humans, which is anomalous from the normal three classes, is inherited through genetics, specifically in the possession of two X chromosomes. Results from genotyping and phenotyping suggest that some females who have been identified as potential tetrachromats may actually have four functional photopigment classes in their retina [62] due to the presence of two variants of L-cone cell pigment genes, one on each X chromosome.

In animals, the existence of four cone cells, in contrast to humans' possession of three, allows them to perceive a wider range of the electromagnetic spectrum. While humans can typically detect light with wavelengths extending from 400–700 nm, tetrachromat animals can perceive light into the ultraviolet range (300–400 nm) [19, 41]. The implications of this phenomenon in other species may suggest that human potential tetrachromats with the right neural programming are capable of a dimension of perceptual hue that may not be experienced by individuals who have only three functional cone classes [62]. This extra cone class is thought to be sensitive to long wavelengths, essentially creating an additional L-cone class. Just as with normal trichromats, potential tetrachromats are posited to exhibit skewed proportions to certain cone class types relative to others [62]. These results are consistent with previous results that potential tetrachromats experience a richer color experience than the more common trichromats [57]. This richer color experience was reported in results from a potential tetrachromat subject in Jameson, Winkler, and Goldfarb's experiment [61] who exhibited: (i) non-deficient and non-normative color perception, (ii) minimum motion of some color stimuli at luminance levels outside of a trichromat's range of isoluminant settings, (iii) greatest deviations from normal observers in the red region of the visible light spectrum, and (iv) enriched color sensations in dim light.

Since the prospect of tetrachromacy among humans is not easily proven empirically, it is expected that this X-chromosome-linked variant from the more common color vision phenotype adds further variation to the possible configurations found in normal color vision individuals. It also contributes in demonstrating how "normal" color perception is an individualized construction that is dependent on biology and experience. There is still, however, controversy and discussion regarding whether these potential tetrachromats have significantly different color vision from normal trichromats. In this investigation, we use minimized flicker settings as a way to compare the perception of trichromat, dichromat, and potential tetrachromat observers.

5.1.3 Minimized Flicker Settings

The goal of the *minimized flicker setting* (also called a critical flicker fusion threshold in other literature), is to establish the minimum frequency at which two physically distinct stimuli can be displayed as an alternating cycle and continue to appear as a stable, homogenous light to the observer. We call this alternating presentation of two stimuli a *flicker*. Thus, the minimized flicker setting is defined to be the presentation speed (in frames per second) at which the appearance of a flicker between two sequentially presented spectra is minimized. A flicker is considered minimized when, gradually increasing the presentation speed, the subject reaches a speed at which the flicker is no longer detectable. Similarly, a flicker is also considered to be minimized when, gradually decreasing the presentation speed, the observer reaches a speed slow enough to detect a flicker between the two displayed spectral stimuli. Factors that influence the determination of this minimized flicker setting across various observers include physiological factors like age or fatigue, the stimulus viewing angle, the duration of exposure to the background light, the luminance level of stimuli presented, and stimuli wavelength ranges.

This investigation hypothesizes that variation in the number of cone classes individuals expressed in conjunction with individual variation in photoreceptor class peak sensitivities may contribute to individual differences in minimized flicker settings across different color vision phenotypes.

5.1.4 Methods for Determining Minimized Flicker Settings

Minimized flicker settings have been used extensively in clinical research over the past several decades, which has consequently increased the demand for standard testing conditions when administering this task [26]. Due to the variability and sensitivity of the minimized flicker setting, many different factors can affect an observer's determination of their minimized flicker setting, allowing this task to serve as a useful measure for various phenomena.

Minimimized flicker settings are known to be affected by factors such as target luminance, target color, and target size [11]. Bharathi and Pothi Reddy created a testing device that incorporated all of these factors under the hypothesis that brighter stimuli will better indicate the presence of cataracts and other media opacities. This device utilized the sound card of the computer and linked it to a square wave digital output between the range of 10–50 Hz in order to display a flickering stimulus through an LED driving circuit. The study concluded that flicker fusion thresholds have a direct relationship with stimulus brightness and varies based on the location of the retina where the perceived image is received.

Additional studies have also been conducted to assess the impact that a flickering stimulus has on humans, using pupil size as the measurement system [89]. Given a standardized measurement system that relates pupil size to color, frequency, and luminance of the flickering lights, it is possible to determine minimized flicker settings and quantify their effect on humans through pupil size measurements.

The study described in this paper utilized an algorithm and user interface, which were linked to a digital micro-mirror device that displayed various stimuli to observers. The program was designed and implemented to determine this minimized flicker setting for all types of individuals assessed.

5.1.5 Applications of Minimized Flicker Settings

The psychophysical task of determining minimized flicker settings has been used widely in previous studies in order to make inferences about health as well as human behavior.

One such application of minimized flicker settings has been to supplement various forms of vision testing. It was found that those with macular degenration reported slower minimized flicker settings than those with healthy eyes [135]. Similarly, the presence of diminished minimized flicker settings was used to diagnose optical hypertensive patients and those with glaucoma [1]. The implication for those who exhibit slower minimized flicker settings is a reduced central sensitivity in the eye to a flickering sensation. In fact, the tests which utilize minimized flicker settings were found to do better in identifying possible early detectors of glaucoma and ocular hypertension than the Snellen acuity test [1]. The results from these studies are consistent with previous findings by Vianya-Estopá that temporal contrast sensitivity measurements, though independent of lens opacities, are very useful in detecting retinal diseases [135].

Additionally, minimized flicker settings can be used as a means to measure intelligence, as drug study indicators, a measure of fatigue, and as diagnostic indicators for some brain diseases. Minimized flicker settings have also been hypothesized to be directly related to perceptual motion learning [118]. Hence, a connection can be drawn between the human brain's ability to better distinguish two lights at quicker amplitudes of modulation and the human body's ability to improve in discriminating coherent motion directions. Another investigation

found that the correlation between an individual's sociability level and a caffeine-induced arousal can be measured by changes in minimized flicker settings [24]. In Corr et al.'s study, subjects first determined an initial minimized flicker setting before receiving a stimulant, either 500mg of caffeine or a placebo, then determined a second minimized flicker setting. They concluded that the magnitude of the change in the minimized flicker settings corresponded to the level of arousal that an individual experienced as a result of either the caffeine or the placebo. Therefore, the authors claimed that the use of change in minimized flicker settings was a valid method of testing Eysenck's theory of personality which then extends to many other arousal-mediated performance paradigms [24]. It is evident from these various studies that flicker fusion thresholds are not necessarily fixed values unique to an individual, but rather are indicators of the mental and physical capacities of that individual at that time.

The desired application of the minimized flicker setting task in this study is to understand differences in perception across observers with varying color vision phenotypes. More specifically, we seek to address the implications of a heterogeneous population of observers living in a world designed using a standardized trichromatic model. The minimized flicker task will allow us to identify subtle differences in perception between individuals, which could help inform future standardized models used in lighting and color industries.

5.1.6 CIE Standard Observer Model

The Commission Internationale de l'Eclairage (CIE) (translated International Commission on Illumination) is an international organization which establishes standards related to measuring and quantifying light and color. Since its inception in 1913, it has become the international authority on matters regarding light, illumination, color, and color spaces.

In 1931, the CIE established a set of functions, called color matching functions, which collec-

tively defined the 2° *standard observer*. The standard observer was designed to represent the chromatic response of an average human using a 2° viewing angle. This particular viewing angle was chosen because, at the time, it was believed that all of the color-sensitive cones were situated within a 2° arc of the fovea. The color matching functions of the standard observer were derived from data collected by John Guild and David Wright in 1927 [29]. In the Guild and Wright experiment, subjects looked through a circular bipartite field that subtended a 2° field of view and were asked to match target colors by combining different intensities of red, green, and blue lights. The responses of subjects (i.e. how much red, green, and blue intensity was used to match each target stimulus) were averaged, creating three color matching functions—$\overline{x}(\lambda)$, $\overline{y}(\lambda)$, $\overline{z}(\lambda)$—which, given any wavelength λ, would reveal how much of each primary color would be needed to match the color of light characterized by that wavelength. The CIE published these functions as the 2° standard observer with the intention of using them to standardize the evaluation of color in industry.

However, in the 1960s, scientists discovered that the field of view of the human eye was actually larger than they had initially thought. This revelation supported the observation that the numerical values calculated using the 2° standard observer functions did not align well with what the actual human eye was perceiving with regards to color. As a result, the CIE published a 10° supplementary standard observer in 1964, using data from the color matching experiment of Stiles and Burch [129] which utilized a 10° field of view. The 10° standard observer is considered to be a better representation of the human eye's visual assessment of color.

In practice, the CIE standard observers are useful when quantitatively evaluating an object's color because they provide a way to translate values taken from color measurement instruments to the the visual experience of actual human observers. Some applications and uses of the standard observer can be found in spectrometer and colorimeter measurements, quality control in lighting, food, and printing industries, as well as other color evaluation

procedures.

5.1.7 Lighting

Of the numerous factors that affect the way in which humans perceive color, lighting is among the most important. The color of an illuminant will determine what color our brain perceives objects under that illuminant as. Though changes in lighting might seem like a trivial problem, multiple studies have demonstrated the affects that lighting can have on the mood and behavior of people [20, 3, 47, 134, 22, 42, 2, 97]. Differences in lighting can be caused by both natural (e.g. changes in amount of sunlight throughout the day) and artificial settings (e.g. rooms with incandescent, LED, halogen, or fluorescent bulbs).

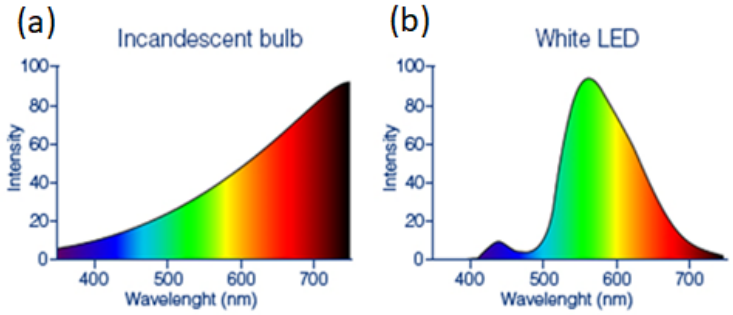

Figure 5.2: Example spectral power distributions of (a) an incandescent light bulb and (b) an LED light bulb. The colors under the curves represent the hue associated with each wavelength within the visible light range (400–700 nm), represented on the x-axis. The values on the y-axis can be interpreted as relative intensities.

With the increasing number of artificial light options, an important consideration in the manufacturing of these lights is how well they can reproduce the colors that we perceive under natural conditions. The Color Rendering Index (CRI) is a measure that quanitatively describes the ability of an artificial, white light source to faithfully reproduce the colors of objects as they would appear under a natural light source. This means that objects illuminanted under a high CRI light source would be perceived as very similar in color appearance as when illuminanted under direct sunlight.

Even lights with very similar properties (e.g. color temperature, a measure of the color appearance of a light source) could have very different color rendering indices. For example, light-emitting diode (LED) light bulbs have very low energy output in the red region in comparison to incandescent light bulbs (see Figure 5.2). Due to this absence of red in the spectral composition of an LED bulb, certain objects—such as those that we perceive as red under natural light—may appear less vibrant in color under an LED light (see Figure 5.3.) Incandescent lights, on the other hand, have energy output across the whole visible spectrum and, thus, have very good color rendering capabilities.

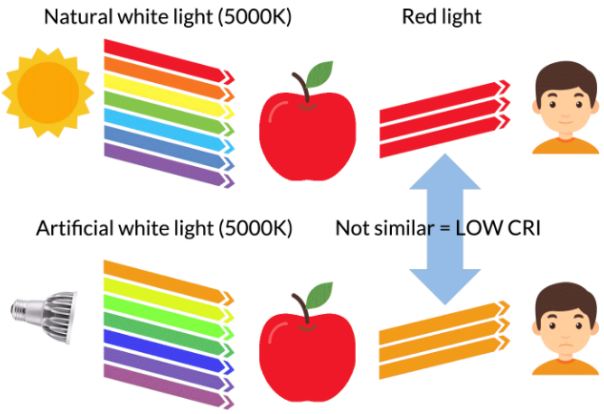

Figure 5.3: Effects of changing illuminants on the colors our brains perceive certain objects as. For example, LEDs have low energy output in the "red" wavelengths, so a red apple might appear less vibrantly red than it would under sunlight. This inability of the LED light to reproduce the same colors we perceive in natural light indicates that it has a low color rendering index (CRI).

Why is this discussion regarding the color rendering of incandescent and LED bulbs relevant? Though incandescent bulbs are preferable because of their ability to accurately reproduce the same colors we perceive naturally, they are very environmentally costly as their high energy output in the red region causes them to require large amounts of energy to function. From an efficiency perspective, it is more advantageous to use LED bulbs because of their reduced energy output in the red wavelengths. However, from a color fidelity perspective, it is more advantageous to use incandescent bulbs because of their higher CRI. An ideal solution would be to develop LED lights with high color fidelity. This study is serving as a preliminary step in understanding how this might be achieved. Another goal of this study

is to determine how these differences in lighting might be affecting the color perception of observers of varying color vision phenotypes.

5.2 Methods

5.2.1 Apparatuses

5.2.1.1 OL-490 Agile Light Source

The digital micro-mirror device (DMD) used to produce spectra for this project was a Gooch and Housego's OL490 Agile Light Source (see Figure 5.4) that utilizes Digital Light Projection (DLP) technology developed by Texas Instruments. DMDs can be used to create spectral profiles at various intensities that are emitted either at a single wavelength or over a broad spectrum. The OL490 Agile Light Source uses a programmable interface that provides the user a large amount of flexibility in producing high resolution spectral output. The control software gives the user control over numerous spectral properties, including wavelength, bandwidth, and intensity, thus giving the user the ability to create a large variety of spectra, combination of spectral lines, and sequences through setting multiple bands, sweeps, and trigger modes. The spectra produced using the control software is emitted using a 3 mm liquid light guide. A 500W Xenon lamp is used to generate output intensities spanning across the visible spectral range.

5.2.1.2 Newport Corportation Integrating Sphere

Stimuli generated by the OL-490 were projected through its liquid light guide and into an integrating sphere. The integrating sphere used in this set-up was a 4 port, 4 inch diameter barium sulfate integrating sphere manufactured by the Newport Corporation. During exper-

Figure 5.4: A visual of (a) Gooch and Housego's OL-490 Agile Light Source and (b) a diagram of the mechanics inside the machine.

imental sessions one of the ports of the integrating sphere was left open, allowing subjects to view inside the integrating sphere. We chose to use an integrating sphere in our experimental set-up because of its properties as a Lambertian reflector—the apparent brightness of a Lambertian surface is uniform independent of the observer's viewing angle [50]. Using a Lambertian surface preserved the uniformity of the stimuli across each of the subjects tested.

5.2.1.3 OceanOptics Flame Spectrometer

An Ocean Optics Flame spectrometer was used for the purposes of calibrating the OL-490 and for measuring spectra displayed by the OL-490. The Flame spectrometer connects to a laptop or desktop PC through a USB port, which then allows the spectrometer to interface with the user-directed OceanView program or the Ocean Optics software development kit. To calibrate the OL-490, one end of a fiber optic cable was inserted into the spectrometer and the other end was inserted into a Newport Corporation integrating sphere. Using this configuration, we took objective measurements of the light being emitted by the ALS. Measurements taken from the spectrometer were compared against the information displayed on the interface of the OL-490 to determine the accuracy of the emissions coming from the light source.

The OceanView program and Ocean Optics software development kit both have the functionality to not only retrieve the spectral power distributions of various spectra, but also

Subject ID	Sex	Age	Color Vision Phenotype	# Sessions
01	F	24	Trichromat	2
02	F	56	Tetrachromat	2
03	F	32	Tetrachromat?	2
04	M	28	Trichromat	2
05	M	55	Trichromat	1
06	F	40	Trichromat	1
07	M	30	Dichromat	1
09	M	24	Trichromat	2
10	M	22	—	1
11	F	30	Tetrachromat	2
12	F	21	—	1
13	F	23	—	1
14	F	22	—	1

Table 5.1: Table of participants.

to compute photometric and colorimetric values. We utilize these features specifically to compute luminance values (in cd/m^2) and the tristimulus values (X, Y, Z).

5.2.2 Participants

A total of 14 individuals (6 males, 8 females; ages 21 to 56) volunteered to participate in this study. Of the 14 participants, 6 (2 males, 4 females; ages 24 to 56) of them were asked to return for a second follow-up task. Prior to testing, a majority of subjects were asked to provide DNA samples so that photopigment opsin genotyping could be performed to test for potential genetic mutations in the subject's color vision. All genotyping investigations were performed with the participants' informed, written consent. Procedures adhered to protocols based upon the world medical association declaration of Helsinki ethical principles for research involving human subjects, and were approved by the ethical review board of the University of California, Irvine. Table 5.1 provides a summary of the participants.

5.2.3 Lab Set-up, Stimuli, and Surround

All stimuli were presented using Gooch and Housego's OL-490 Agile Light Source. Subjects were seated at a booth facing a panel with a 2.75 inch diameter viewing hole. The OL-490 was situated on the opposite side of a black barrier. The OL-490's liquid light guide was inserted into a 1 inch port of a 4 inch integrating sphere. A 1.5 inch port that was facing the subjects was left open to allow subjects to view into the integrating sphere (see Figure 5.5 for visuals of the experimental set-up).

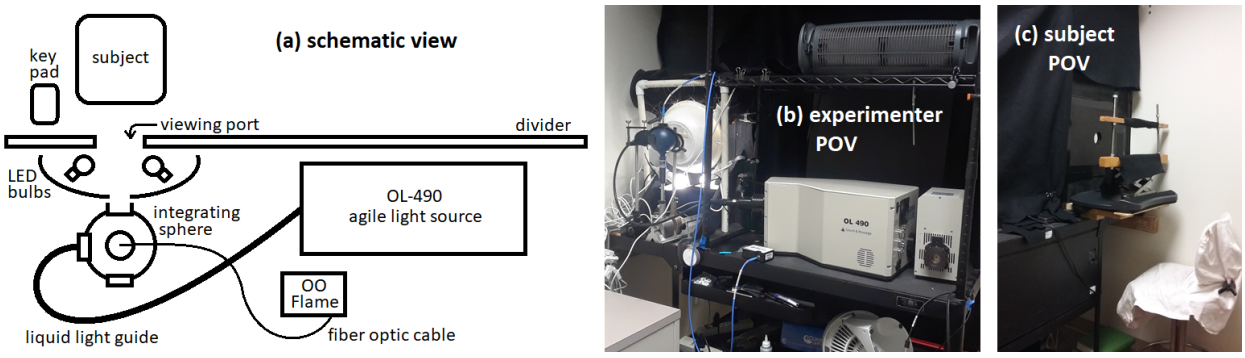

Figure 5.5: A (a) schematic view of the experimental set-up in the laboratory with accompanying photos of (b) the testing booth from the experimenter's point of view and (c) the subject's point of view.

The viewing hole was larger than the open port of the integrating sphere, creating an annulus surround around the target stimuli. The annulus surround was illuminated using 4 LED light bulbs. The surround was achromatic with a measured brightness of ~ 10 cd/m^2. The achromatic annulus provided a visual context was a minimal relational stimulus configuration, as opposed to the standard Maxwellian viewing configuration. Subjects were positioned approximately 10 inches away from the surround in order to achieve the correct viewing angles, which was measured to be 6.439° with respect to the center of the viewing port and 8.578° to 14.25° from the inner edge to the outer edge of the achromatic annulus.

The stimuli we used were designed specifically for this study by Lorne Whitehead from the University of British Columbia.[1] For the first task of the study, we matched a standard

[1]Dr. Whitehead is an applied physicist who has contributed to the fields of optics and illumination

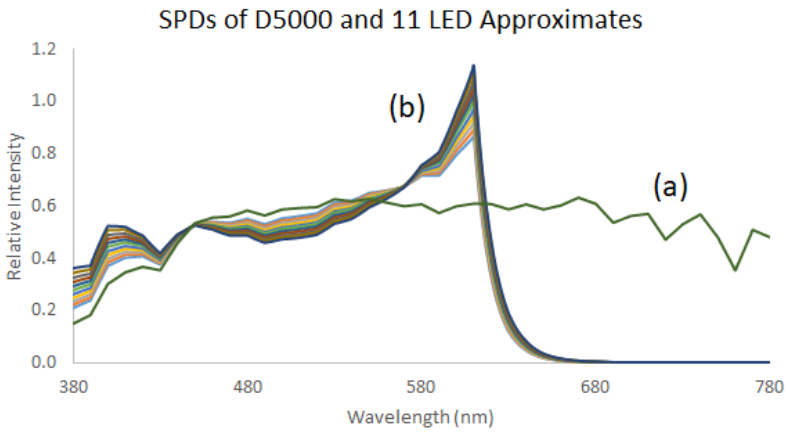

Figure 5.6: Spectral power distributions of (a) the D5000 illuminant and (b) 11 truncated spectra that were designed to match or closely match the D5000 in chromaticity. The 11 truncated spectra were varied in hue along the red-green axis of the color space such that 1 spectra was constructed to match all the tristimulus values of the D5000, 5 spectra were made incrementally more red in color appearance, and 5 spectra were made incrementally more green in color appearance.

illuminant with a color temperature of 5000K, called the D5000, with 11 truncated spectra varied along the red-green axis (see Figure 5.6). The truncated spectra were designed to mimic the spectral power distribution of LED bulbs (i.e. lacking energy in the "red" wavelengths; for this reason, these spectra will be referred to as "truncated approximates" and "LED approximates" interchangeably) and have the same chromaticity as the D5000. One of the 11 truncated spectra was modeled using the three chromatic response curves of the 10° standard observer. Five of the other truncated spectra were varied to appear more red in hue and five others were varied to appear more green in hue. For the second task of the study, we were interested in the effects of these different illuminants on individuals' ability to render and perceive the color of objects, so we convolved the D5000 and the 11 truncated spectra with the RHEM light indicator. The RHEM light indicator is an indicator composed of two spectra that appear metameric (i.e. same in color appearance, but with different spectral power distributions) only under a standard 5000K illuminant (see Figure 5.7).

through the development of patented tools and license agreements. He is also a member of the International Commission on Illumination (CIE).

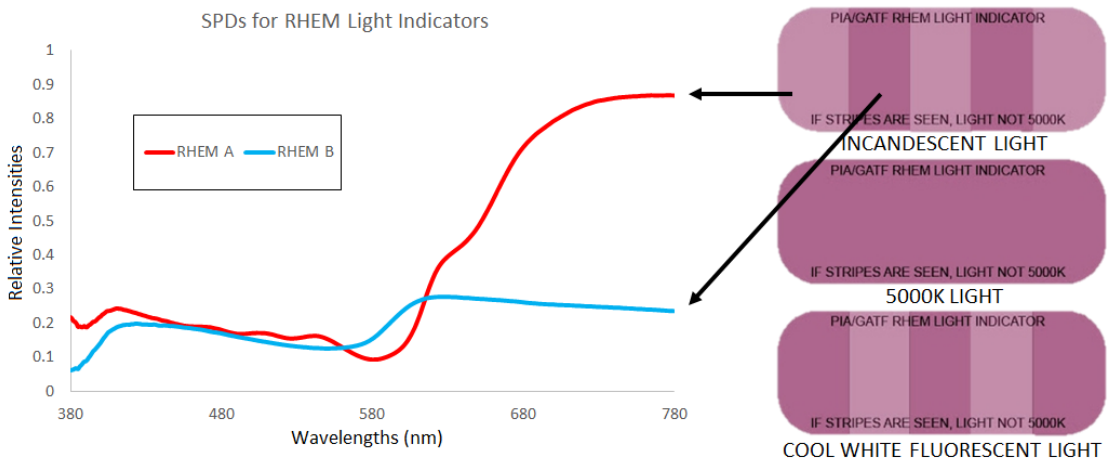

Figure 5.7: Spectral power distributions of the two strips on the RHEM light indicator.

Stimuli were presented to subjects as flickering pairs. Experimental pairs consisted of two spectra that were fundamentally different from each other—that is, they had distinct spectral power distributions. Foil pairs consisted of two identical spectra. Foil pairs were included as a way to test subjects' understanding of the task. During the presentation of a foil pair, subjects should not detect any flicker because there is no physical change to be detected due to the fact that the spectra are identical. If a subject claims to detect a flicker during a foil trial, it can be presumed that the subject most likely did not understand the nature of the task and can therefore be considered an unreliable subject. The spectral pairs were presented sequentially in a randomized order. For each spectral pair, subjects were required to manipulate the flicker speed until they found their minimized flicker rate.

5.2.4 Procedure

Prior to each session, subjects were given a 10 minute adaptation time to adjust to the darkness of the room. The experiment was completed in two sessions (some subjects only participated in one session and some participated in both). These sessions were usually completed on separate days, with the exception of one subject who had to complete both

sessions in a single day due to time constraints. The first session lasted about 30 minutes and the second session lasted anywhere from 1.5 to 2 hours. Each session was divided into two sub-tasks, which we will refer to as the two staircase types.

5.2.4.1 Staircase Types

In each of the stimulus conditions, Illuminant and RHEM-Illuminant (see Section 5.2.4.2), subjects completed two sub-tasks called the *observer-directed* and *adaptive* staircases. The observer-directed staircase has been used in previous pilot and demo studies by Joe and Jameson and was determined to be a useful, reliable task for determining minimized flicker settings. However, the adaptive staircase is a more commonly used method in psychophysical and signal detection tasks [75, 130, 120]. Therefore, subjects completed both tasks using both staircase types and the results were compared to see if the two methods yielded significantly different results[2].

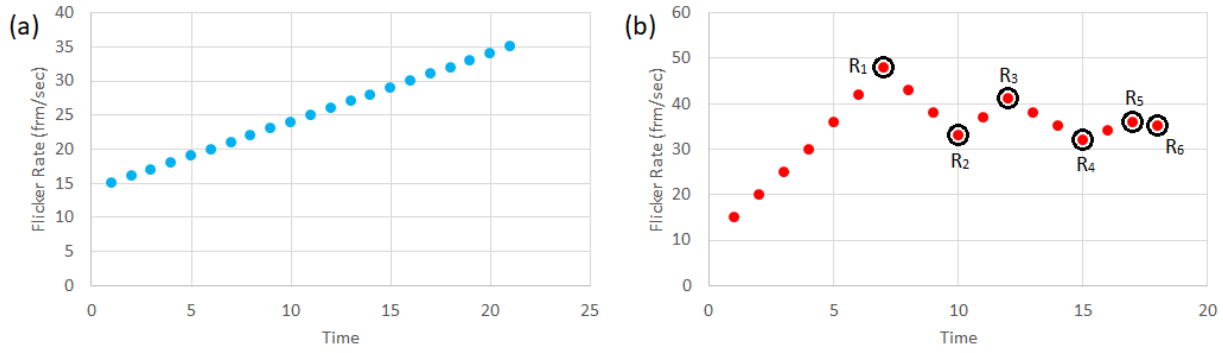

Figure 5.8: Graphical depictions of trials of the same stimulus pair using (a) the observer-directed staircase and (b) the adaptive staircase. Each dot represents a single 1.5 second stimulus presentation, and its y-value represents the flicker rate of that presentation. In the observer-directed staircase, subjects can only change the flicker speed 1 fps at a time and they have the freedom to terminate the trial whenever they decide they have found their minimized flicker setting. In the adaptive staircase, the change in presentation speed from dot to dot is variable and the end of a trial is endogenously determined once the subject has six direction reversals.

[2]This comparison was not able to be done due to unexepected circumstances which prevented this study from being completed. These comparative results will be reported in future work

Observer-Directed Staircase

The observer-directed staircase method allows subjects to directly determine their minimized flicker setting. In this method, subjects manipulated the presentation speed 1 unit (frame per second) at a time and were asked to determine the point at which the flicker became "minimized" (see Section 5.1.3 for definition of a "minimized flicker"). A trial started with a spectral pair being flickered at some initial presentation speed for 1.5 seconds before blacking out. If the subject could detect a flicker in the display, then they were instructed use the "up" key on the keypad to increase the presentation speed until the flicker could no longer be detected. If they could not initially detect a flicker in the display—i.e. the display looked like one solid color—then the subject was instructed use the "down" key on the keypad to decrease the presentation speed until the flicker became discernable[3].

Adaptive Staircase

In the adaptive staircase method, the computer used an adaptive procedure to change the flicker rates, allowing subjects to traverse the space of possible presentation speeds more quickly. With every key press, the computer changed the presentation speed by Δ units, where Δ was varied based on the history of key presses. Subjects did not directly determine their minimized flicker setting for each flickered pair; rather, the computer terminated a trial after 6 staircase reversals (see Figure 5.8). The adaptive procedure used in this study was adapted from the procedure used by Paramei to administer the Cambridge Colour Test [101], which was derived from the work of Regan, Reffin, and Mollon [107, 91].

During this task, the only judgment subjects had to make was whether they could detect a flicker or not. Flickering pairs were presented for 1.5 seconds before the display blacked out. After the pair had been displayed, subjects were instructed to use a keypad to answer

[3]When subjects reached either of these points, they were instructed to press an "enter" key which would communicate to the system that they had found their minimized flicker setting and would initialize the next stimulus pair in the sequence. While decreasing the presentation speed, if subjects ever reached the minimum possible presentation speed (1 frame per second) they were given a feedback noise indicating that there was no lower speed they could reach. Subjects were instructed to press the enter key when they heard this noise to move on to a new pair of stimuli.

the question: "Did I see any flicker in the display?" Pressing the "yes" key would increase the flicker speed by Δ frames per second. Pressing the "no" key would decrease the flicker speed by Δ frames per second. At the beginning of each trial, Δ (i.e. the change in flicker speed as a result of a key press) was set to 5 frames per second (frm/sec). For example, if the initial presentation speed was 15 frm/sec, pressing the "yes" key would change the flicker speed to 20 frm/sec for the following stimulus presentation. If a subject continued in the same direction (i.e. presses the same key) three times in a row, Δ would increase by 1 frm/sec. If a subject changed directions, then Δ would decrease by 1 frm/sec. The smallest Δ could ever become was 1 frm/sec. The trial terminated after the subject had made six reversals. When a trial terminated, a new stimulus pair was presented to the subject and Δ was reset to 5 frm/sec.

5.2.4.2 Stimulus Conditions

Condition 1: Illuminant Stimuli

Following the 10 minutes of dark adaptation, subjects were given 4 practice trials using the observer-directed staircase method. Then, subjects completed a sequence that consisted of 22 trials (two repetitions of 11 spectral pairs) also using the observer-directed staircase method, where each trial was a spectral pair consisting of the D5000 and one of the truncated approximates. Subjects were given a brief rest period, then repeated the task (including the practice trials) again using the adaptive staircase method with a new randomized sequence of trials. Average time to complete the task was about 30 minutes.

Condition 2: RHEM-Illuminant Convolved Stimuli

Several subjects who participated in Task 1 were asked to return on a separate day to participate in this follow-up task. As in the first task, subjects were dark adapted for 10 minutes. A coin was flipped to determine which staircase method the subject would use first. Using the staircase method determined by the flip, subjects completed 6 practice trials (4

experimental, 2 foil) before completing a randomized sequence of 78 trials (70 experimental, 8 foil). The different combinations of convolved stimuli to create spectral pairs for the trials are detailed in the Results section (Section 5.3.2). Subjects were given a brief rest period, then repeated the task again using the other staircase method with a new randomized sequence of trials. Average time to complete the task was about 1.5 hours.

5.2.5 Software Development

Mathematica was used to write the code that facilitated the experiment. The code performed the functions of connecting the program to the OL-490 ALS, generating the stimuli sequence files, displaying color pairs on the OL-490 according to that sequence file, and recording observer responses to each displayed color pair. All of the code that was written to handle controlling the agile light source and using it to display spectra was taken from the OL-490 Software Development Kit v1.3 as well as functions written by Tim Satalich, which handled interacting with the OL-490, calculating corrected inputs to the OL-490, and measuring spectra.

5.3 Results

5.3.1 Stimulus Condition 1: Standard and Truncated Illuminants

Based on the performance of the 14 subjects[4] who participated in Stimulus Condition 1, we can broadly partition them into three groups:

1. Those who did the task reliably (i.e. small standard deviation over repeated trials, as

[4]All of the results reported below have been gathered using data from the observer-directed task. Results from adaptive staircase procedure are forthcoming.

depicted through the error bars) and exhibited the expected "V"-shaped curve.

2. Those who did the task reliably but failed to find a truncated SPD that best-matched the D5000.

3. Those who could not do the task reliably.

Group 1: Reliable subjects with best-matched LED approximate

Out of 14 total subjects, 6 of them were identified as reliable and were able to successfully find a truncated spectrum that was a perceptual best match to the D5000. The best match can be determined visually by finding the dip in the curve that creates the "V" shape (see graphs in Figure 5.9). For example, Subject 01's best match to the D5000 was the 102% X-value deviated truncated spectrum. Figure 5.10 is a histogram which shows the distribution of best matched truncated spectra to the D5000 for the six subjects who were able to find a best match.

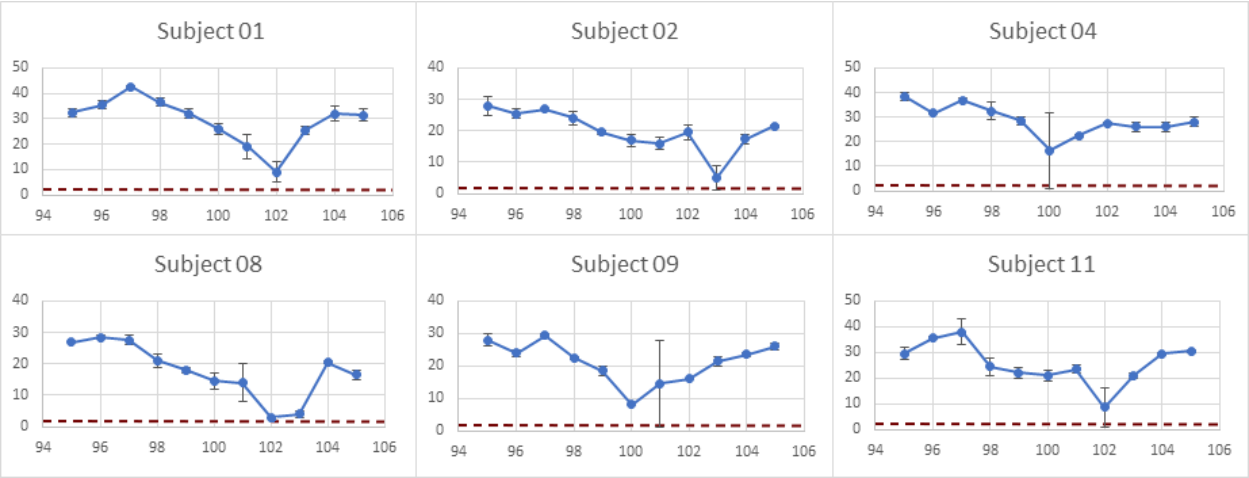

Figure 5.9: Minimized flicker settings across the 11 spectral pairs displayed for the 6 subjects in Group 1. The x-axis represents the deviation of the X value of the truncated spectra, ranging from 95% (more green) to 105% (more red) with 100% representing the truncated SPD designed to match to the D5000 in chromaticity, according to the CIE 10° standard observer. The y-axis represents the average minimized flicker setting in frames per sec (frm/sec) for the repeated trials of a specific spectral pair. Error bars represent one standard deviation from the mean. The dashed line represents the "floor" or the lower bound of possible minimized flicker settings which is equated to 1 frm/sec.

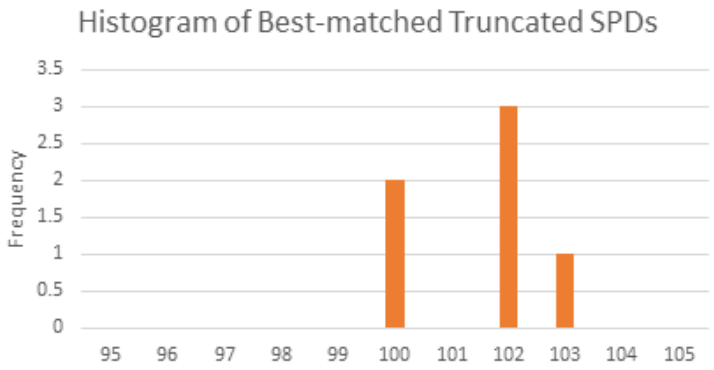

Figure 5.10: Histogram of best-matched truncated spectra to the D5000.

The first interesting observation is that 4 out of the 6 subjects in this group found a best match that was different than the best match modeled by the 10° standard observer. This implies that the predictions being made by the CIE standard observer model may not be a suitable representation of how individuals are actually perceiving color. Therefore, subjects' perception appears to be deviated from the predictions made by the standardized model.

In addition to the observed overall deviation from the standard observer model, the direction of the deviation appears to be systematic across subjects. All six subjects found their best matched LED approximate to be equal to the 10° standard observer best match or deviated towards the red region of the red-green axis (i.e. all truncated spectra that were found to be a best match to the D5000 had a deviated X-value of 100 to 105%). This finding is consistent with previous reports that stimuli designed to match target stimuli using the CIE standard observer models appeared more green-ish in hue than their target counterparts. In this case, with regards to the D5000 (which appears achromatic), we expected individuals to find their best matched truncated D5000 approximates in the red direction in order to cancel out the green color apparent in the original (100%) truncated spectrum. Though we would expect some divergence from the standard model, the fact that subjects are deviating from the standard model in the same direction suggests the presence of a systematic bias rather than natural variation around an average.

Group 2: Reliable subjects with no best-match

Two subjects were identified as reliable, due to their relatively small standard deviations across the repeated trials for each spectral pair, but differed from those in *Group 1* in that their results exhibited a more linear trend as opposed to the expected "V"-shaped curve. This pattern suggests that all of the 11 truncated spectra were all equally ill-fitting or equally different to the original D5000. That is, for these subjects none of these truncated spectra were a close perceptual match to the D5000. Since these subjects were found to be consistent in setting similar minimized flicker settings for the same spectral pair across repeated trials, it would be difficult to make a case that these subjects exhibited a different pattern due to lack of understanding of the task. Rather, a more likely reason for this difference in pattern can be attributed simply to the dimensionality of the color space. The 11 truncated spectra used in this experiment were sampled from along the red-green axis of the color space. However, color spaces are often modeled as three-dimensional spaces, so it is possible that the theoretical truncated spectra that is a perceptual best match to the D5000 for these subjects lies off of the one-dimensional red-green axis. This result could be further explored by testing a wider sample of truncated spectra, by perhaps sampling from an ellipse or an ellipsoid in the color space instead of a line.

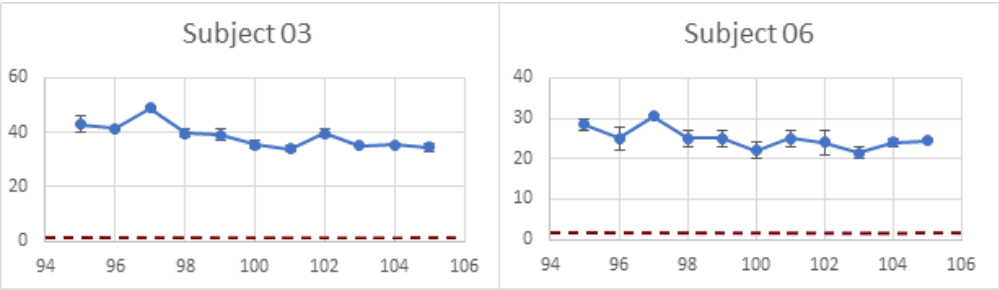

Figure 5.11: Minimized flicker settings across the 11 spectral pairs displayed for the 6 subjects in Group 2. See Figure 5.9 for explanation of the axes and other features of the graphs.

Group 3: Unreliable subjects

Several subjects were flagged as unreliable due to the apparent lack of any systematic trend across the different spectral pairs and/or the presence large standard deviations, indicating

that they were not making consistent judgments across repeated trials. These subjects were omitted from any further data analysis because their results suggest that they either did not understand the nature of the task or where physically incapable of completing it. This task has been shown to be difficult to complete due in part to the subtle nature of the flicker phenomenon close to or at its fusion threshold. Therefore, the ability to do this task well requires a somewhat extensive amount of training and practice. All of the subjects in this group, except for one, had not previously participated in an experiment using this minimized flicker task. Therefore, it is likely that many of these subjects did not have an adequate amount of experience with the task to give reliable results.

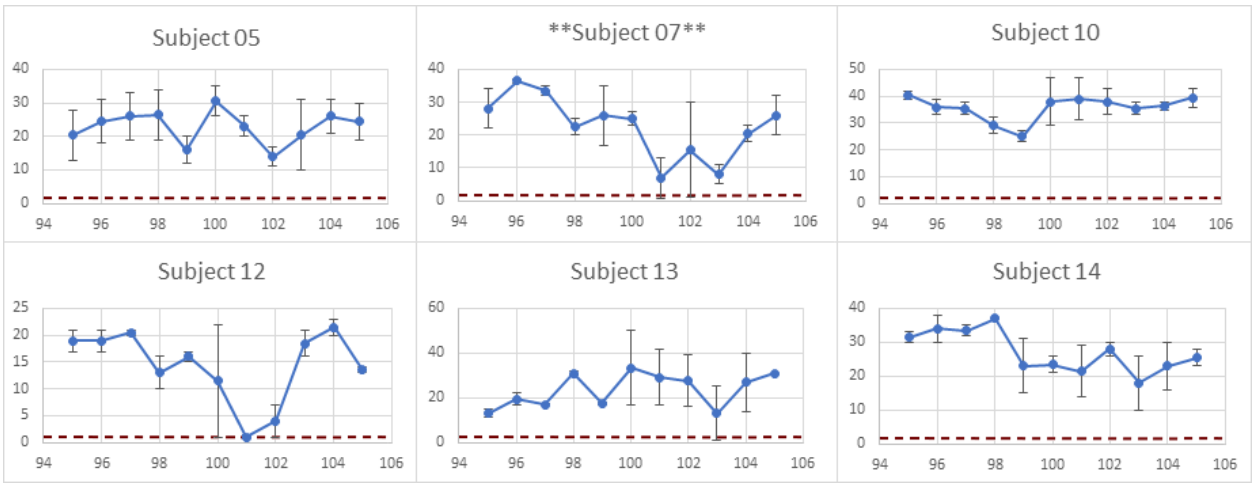

Figure 5.12: Minimized flicker settings across the 11 spectral pairs displayed for the 6 subjects in Group 3. The dashed line represents the "floor" or the lower bound of possible minimized flicker settings which is equated to 1 frm/sec. See Figure 5.9 for explanation of the axes.

Subject 07 has been starred because he is a dichromat (deuteranope). He has participated in previous studies which utilized this same minimized flicker task on a different set of stimuli, so his inconsistency in performance is probably more a function of his color vision phenotype as opposed to a lack of experience or understanding. In fact, his inconsistency is expected since the truncated spectra were all deviated along the red-green axis and his particular deficiency is in his M-cones (colloquially known as red-green colorblindness). Since his color vision phenotype prevents him from perceiving red and green hues, we would expect his results

to be unsystematic across the various spectral pairs. Subject 07's performance reveals the limitations of the task and shows that the task fails in situations where we would expect it to fail.

A quantitative criterion for determining whether certain data should be omitted is in the process of being developed.

Aggregate results across reliable subjects

Simple hypothesis tests were conducted on all the reliable subjects' (*Groups 1* and *2*) average minimized flicker settings for the D5000 and CIE observer best-match (truncated approximate at 100%) to see if they were statistically significant from the "floor" (i.e. 1 frm/sec, the lowest possible presentation speed). A minimized flicker setting of 1 frm/sec indicated that subjects could not detect any difference between the two stimuli being presented even at the lowest possible presentation speed. An alternative interpretation of a minimized flicker setting of 1 frm/sec is that the two stimuli appeared metameric to the observer. Any minimized flicker setting greater than 1 indicated that the two stimuli being presented did not appear identical to the observer.

All subjects except for one[5] had an average minimized flicker setting that was statistically greater than 1 frm/sec at the 0.05 level (see Table 5.2).

5.3.2 Stimulus Condition 2: RHEM-Illuminant Convolved Stimuli

From the 8 reliable subjects in Group 1, 6 of them were asked to return and participate in Stimulus Condition 2. This stimuli used in this condition were convolutions of the RHEM indicator strips with the D5000 and a subset of the truncated approximates from Stimulus

[5]Subject 04 had a very large standard error for this particular spectral pair because he made an error on one of his trials. Since the number of repeated trials in this task was very low, this one error caused the standard error to become very large. His result probably would have changed if he had done more repeated trials.

Subject ID	Avg MFS for (D5000//DT_{100})	P-value	Metameric?
01	26	0.017987	No
02	17	0.028062	No
03	35.5	0.009783	No
04	16.5	0.195913	Yes
06	22	0.021404	No
08	14.5	0.041445	No
09	8	0	No
11	21	0.022471	No

Table 5.2: Results from the hypothesis test $H_0 : \mu_{MFS} = 1, H_1 : \mu_{MFS} > 1$ for the 8 reliable subjects from groups 1 and 2. Based on the results from the hypothesis test, we can infer whether the participant perceived the D5000 and the truncated approximate modeled using the CIE standard observer (DT_{100}) as metameric. The last column of the table details whether the participant perceived the two spectra as metamers or not.

Condition 1. This condition was designed to be more psychophysically robust than the first condition and therefore was comprised of more repetitions and foil trials. As mentioned in Section 5.2.3, foil trials were comprised of two identical stimuli and were included for the purpose of checking the subjects' understanding of the task. All subjects successfully completed the foil trials, which means that no one detected a flicker phenomenon when none was present. Therefore, this provides some support for the reliability and trustworthiness of these subjects' performance.

The plots in Figure 5.13 detail the results from 6 subjects who participated in this task. The notation for the different spectral pairs depicted on the x-axis is as follows:

- An "x" represents the convolution of two spectra.
- "A" represents RHEM A and "B" represents RHEM B.
- "D5000" represents the standard D5000 illuminant.
- "DT" represents the truncated approximate of the D5000 modeled using the CIE 10° standard observer.

- "DT*" represents the truncated approximate of the D5000 which the subject identified as their best match to the D5000 in Stimulus Condition 1.
- The two spectra that make up the pair in the flicker are listed along the x-axis. The pair is comprised of the spectra that are vertically adjacent to one another.

Subject 03 was one of the subjects from Task 1 who did not find a best-matched truncated SPD. Since she did not have a DT*, we set DT* = DT when generating the randomized sequence of stimuli for her experimental session. Subjects 04 and 09 both found their best matches to be the same as the CIE prediction, so they also had DT* = DT. Therefore, the stimuli from pairs which used the DT* illuminant were not distinct from those which used the DT illuminant. For this reason, average minimized flicker settings for Subject 03, 04, and 09 for spectral pairs (DT*xA and DT*xB), (D5000xA and DT*xA), and (D5000xB and DT*xB) have been omitted and have been agglomerated with pairs (DTxA and DTxB), (D5000xA and DTxA), and (D5000xB and DTxB), respectively.

An initial observation which is evident through the graphs in Figure 5.13 is that all minimized flicker settings are significantly different from 1 frm/sec. That is, a metamerism was not established for any of the spectral pairs shown to any of the observers. This is perhaps most noteworthy for the first pair: (D5000xA and D5000xB) (see Table 5.3). By construction, the RHEM stimuli are designed to be metameric under the D5000 standard illuminant. Therefore, we expected that the convolution of the D5000 with the two RHEM strips would result in a pair of metamers. It is apparent from the relatively high minimized flicker settings for the (D5000xA and D5000xB) pair that our hypothesis was incorrect. The failure of the metamerism, however, is not wholly surprising because (i) the intended medium for the RHEM indicator is reflectance on a surface, not emissive lights as we employ them and (ii) the metamerism on the physical indicator strip is weak even in the correct viewing circumstances, so it is possible that the break in the metamerism only gets more dramatized when changing the medium under which the spectra are viewed. In an effort to evaluate

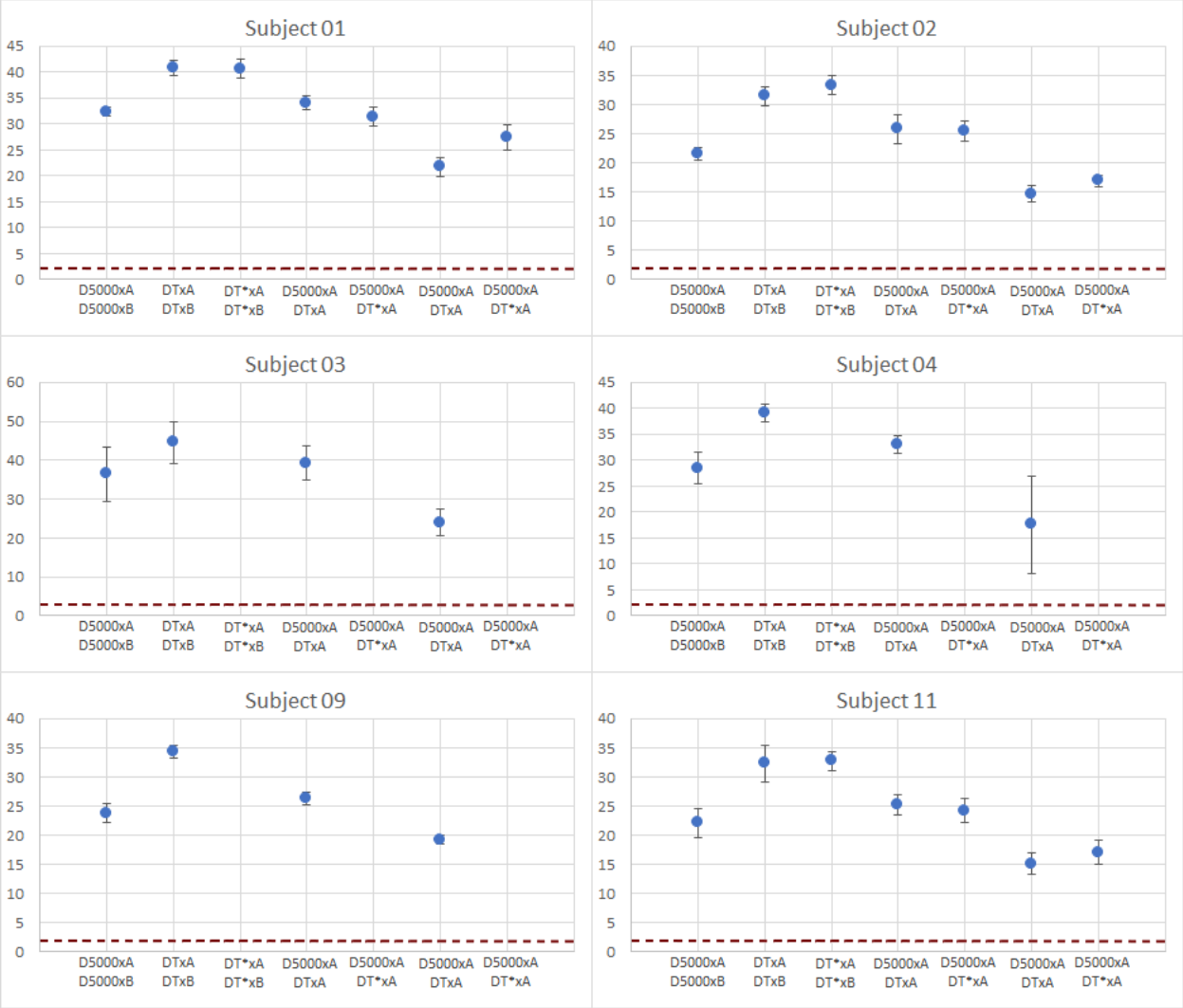

Figure 5.13: Minimized flicker settings across 7 spectral pairs for the 6 subjects in who participated in the RHEM-illuminant stimulus condition. The x-axis represents the different types of spectral pairs. The y-axis represents the average minimized flicker setting from the repeated trials for a specific spectral pair. The dashed line represents the "floor" or the lower bound of possible minimized flicker settings which is equated to 1 frm/sec.

how this metamerism might be affected by changing the medium under which it is viewed, the D5000xA and D5000xB spectra were measured from the integrating sphere and their tristimulus values as well as their LAB and LUV coordinates were computed. Figure 5.14 show the D5000xA and D5000xB stimuli plotted in CIE 1976 color space. When plotted in this color space, the two stimuli plot very close to each other. Their nearness is also exemplified by the small ΔE values between the LAB and LUV coordinates of each stimulus. Though the model predicts that these two stimuli should be nearly indistinguishable from one another, our empirical results suggest that subjects could quite clearly perceive a difference between the two. One possible reason for this incongruity could be a failure of the CIE's standard model to take individual variation into consideration. Hence, the predictions and expectations we have using the standardized models are not matching up with what observers are actually reporting.

Subject ID	**Avg MFS for (D5000xA//D5000xB)**	**P-value**	**Metameric?**
01	32.4	0	No
02	21.5	3.4×10^{-13}	No
03	36.4	2.91×10^{-8}	No
04	28.5	3.4×10^{-10}	No
09	23.8	3.26×10^{-12}	No
11	22.1	4.16×10^{-10}	No

Table 5.3: Results from the hypothesis test $H_0 : \mu_{MFS} = 1, H_1 : \mu_{MFS} > 1$ for the 6 subjects who participated in Stimulus Condition 2. Based on the results from the hypothesis test, we can infer whether the participant perceived RHEM A and RHEM B as metameric when illuminated by the D5000. The last column of the table details whether the participant perceived the two spectra as metamers or not.

A second observation is that the minimized flicker settings for (DTxA and DTxB), (DT*xA and DT*xB) are consistently higher than the minimized flicker settings for (D5000xA and D5000xB) across all subjects. This disparity is confirmed by a series of two sample hypothesis tests which show that the spectral pairs which include a truncated illuminant yield significantly higher minimized flicker settings than the spectral pair comprised of RHEM A and B under the standard illuminant (see Table C.1 in Appendix C for detailed test results).

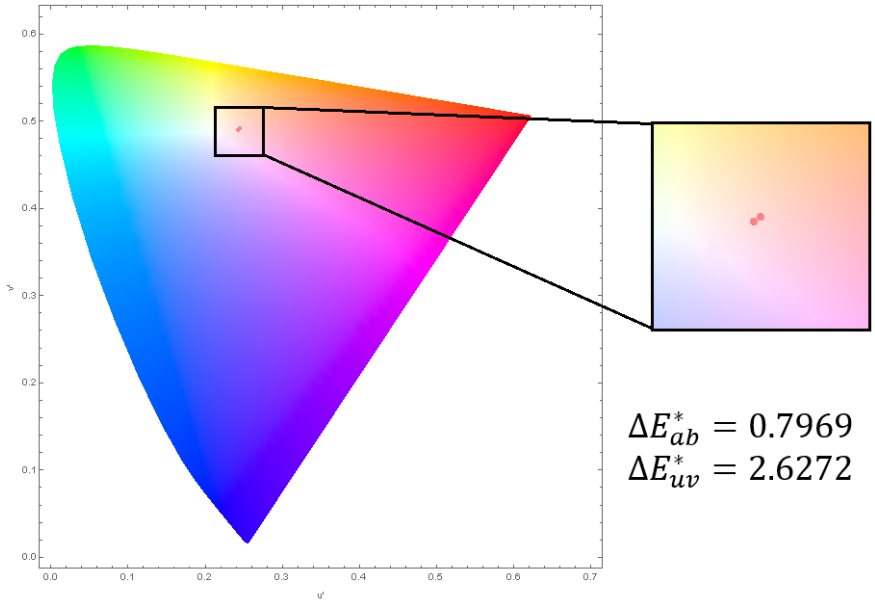

Figure 5.14: D5000xA and D5000xB stimuli plotted in a CIE 1976 u'v' chromaticity diagram. The ΔE values represent the perceptual distance between the two stimuli and are given for both the CIELAB and CIELUV color spaces.

This pattern indicates that the metamerism of the RHEM A and B stimuli is better preserved under the D5000 illuminant than under any of the truncated D5000 approximates. So, even though the metamerism is failing for every pair of stimuli, it appears that the use of an LED illuminant is dramatizing the difference between RHEM A and B. This result is unsurprising given the computed ΔE values in Table C.2 in Appendix C. These values show that the perceived distance, both in CIE LAB and LUV color space, between RHEM A and B is greater when illuminated by one of the LED approximates than by the standard D5000.

Lastly, all of the spectral pairs which were composed of a single RHEM stimulus convolved with two different illuminants (e.g. D5000xA and DTxA) yielded minimized flicker settings that were all significantly different from 1 frm/sec (see Table C.3 in Appendix C). This implies that the same stimulus (one of the RHEM spectra) is being perceived as significantly different when illuminanted by the standard D5000 versus its truncated approximates. Therefore, similarly to the observation listed previously, we can further conclude that the viewing experiences of the observers are being affected by the change in illuminant type, at least in

this particular scenario. We can also use objective ΔE measures to support these results from the experiment. Table C.4 in Appendix C contains ΔE values for pairs of stimuli that represent one of the RHEM spectra being illuminated by either the D5000 or one of its truncated approximates. The ΔE values further reveal the disparity that is being caused by changing from a standard illuminant to an LED approximate.

5.4 Discussion

The primary aim of the illuminant stimulus condition was to evaluate the sufficiency of the CIE 10° standard observer model to represent a heterogeneous population of observers. Preliminary results from this study seem to suggest that the 10° standard observer does not generalize well to individuals who deviate from the standardized model. This finding is consistent with work by Webster and colleagues on individual differences even amongst the population of normal observers [136, 138, 139, 140, 87, 63, 28]. The task used 11 truncated LED approximates of the D5000 illuminant which were varied in their X value along the red-green axis of the color space. Some observers were able to find a single truncated approximate that served as a personalized best match to the D5000 and others were not. Of those who were able to obtain a best match, a (somewhat narrow) majority found their best match to be different than the theoretical best match predicted by the 10° standard observer model. Additionally, for the spectral pair composed of the D5000 and the 10° observer's best match, all subjects (except for one) set their minimized flicker settings significantly higher than 1 frm/sec which indicates that they perceived the the two spectra as distinct. In other words, despite the truncated approximate being modeled by the standard observer to match the D5000 in tristimulus values, almost all of the subjects failed to perceive the two the illuminants as a metameric pair. These observations suggest that the individual variation among observers—even among trichromats—is not being captured by this standard observer

model. Moreover, the truncated spectra which were identified by subjects as their best match to the standard D5000 illuminant were all found to be deviated in the same direction along the red-green axis. This consistency in deviation across subjects suggests the presence of a systematic bias in the standard observer.

The RHEM-illuminant condition's purpose was to evaluate how observers' viewing experiences might change when viewing a natural scene under a truncated LED illuminant versus a full-spectrum standard illuminant. With LED bulbs becoming more popular lighting options due to their energy and environmental efficiency, an important question to ask is how this switch to an illuminant with very different spectral power distributions might be affecting observers' color perception. Therefore, we did not want to just test subjects' perception of the illuminants themselves (as in the illuminant condition), but we also wanted to test how perception of the color appearance of objects was affected by different types of illuminants. This task utilized stimuli from the RHEM light indicator, which is comprised of two spectra that are designed to be metameric under a standard D5000 illuminant. Subjects were presented stimuli that were convolutions of RHEM A or B with the D5000 or one of its truncated approximates. All subjects reported minimized flicker settings that were statistically different from 1 frm/sec for all pairs of spectra, meaning that none of the pairs of spectra were perceived as metameric. Therefore, for pairs which compared the same RHEM stimulus under two different illuminants (e.g. D5000xA and DTxA), subjects perception of the RHEM stimulus was changed when viewing it under a standard D5000 illuminant versus an LED illuminant. Also, the average minimized flicker settings for RHEM A and B under the same truncated illuminant (e.g. DTxA and DTxB) were consistently higher than the average minimized flicker settings for RHEM A and B under the standard illuminant (e.g. D5000xA and D5000xB). This means that even though the truncated illuminants were designed to match the D5000 in chromaticity and color temperature, the LED illuminants are worse at preserving the metamerism between the RHEM A and B stimuli than the standard full-spectrum illuminant. Therefore, these results suggest that use of LED illuminants in

place of incandescent illuminants does affect the color viewing experiences of observers in a non-trivial way.

Among the objectives mentioned above, one of the initial motivations behind the design of this study was to investigate how perception under these artificial light sources differed for individuals with varying color vision genotypes given that these artificial light sources were being designed under a trichromatic model. Since other works [61, 18] have exhibited that the color perception and color processing of human potential tetrachromats differ from that of normal trichromats, a natural question to ask is how potential tetrachromats' color perception is affected when living in a world designed to fit the viewing experience of a trichromat. Due to the small sample size of this experiment, conclusions that can be drawn about differences in perception across different color vision phenotypes are limited and should be taken as preliminary intuitions which will dictate future research. A larger sample is needed in order to draw meaningful inference about differences between the overall trichromat and potential tetrachromat populations.

5.5 Conclusion

The main goal of this study was to serve as a small demonstration which showed ability of minimized flicker settings to identify subtle differences in the color perception across a set of heterogeneous individuals and to indicate the need to rethink the standard observer and its use in industry. This study was intended to serve as motivation for embarking on a larger scale project. The ultimate goal of this line of research would be to extend the standard observer in such a way that it would encompass individual variations and even other color vision phenotypes, such potential tetrachromats. Since this was only designed to be a demonstration, there are many natural extensions to this project which could lead to more precise and robust findings. For example, it could be prudent to test a wider range

of truncated illuminants. Here, we only sampled approximated illuminants from a line in the color space, but sampling from an ellipse or ellipsoid would be a more robust task. Sampling from a higher dimensional space might give us more insight into the directions in which different color vision phenotypes/genotypes vary from each other. Moving forward, we hope to further refine and develop this procedure to ultimately gain a better, more robust understanding about the human color perception and processing (specificially in human potential tetrachromats).

One of the conclusions of this study is that the CIE standard observer model is not a suitable representation of the general population of observers, or even of the population of trichromat observers. Since the current method of averaging is not producing a desirable model that captures the variation in color vision among observers, one possible application of this research could be either determining a more appropriate method for creating a singular representation of the population or extending the model in a way that would be more inclusive of individual variation among observers, including those who express a different number of photopigment classes. Having a better-fitting standard model that could be applied to applications in lighting and color displays could have large implications for the development of more personalized settings and systems—allowing all individuals to achieve the most suitable color environment for themselves.

There are still many more facets of this research that need to be explored. We hope to continue developing studies which investigate possible recommendations for the advancement of the standard observer model and the lighting industry, such as the development of LED bulbs with high color fidelity or the expansion of the CIE standard observer models to encapsulate more of the observed individual variation in color vision.

Bibliography

[1] A. Atkin, I. Bodis-Wollner, S. M. Podos, M. Wolkstein, L. Mylin, and S. Nitzberg. Flicker threshold and pattern vep latency in ocular hypertension and glaucoma. *Investigative ophthalmology & visual science*, 24(11):1524–1528, 1983.

[2] S. Barbut. Effect of illumination source on the appearance of fresh meat cuts. *Meat Science*, 59(2):187–191, 2001.

[3] R. A. Baron, M. S. Rea, and S. G. Daniels. Effects of indoor lighting (illuminance and spectral distribution) on the performance of cognitive tasks and interpersonal behaviors: The potential mediating role of positive affect. *Motivation and emotion*, 16(1):1–33, 1992.

[4] A. Baronchelli, T. Gong, A. Puglisi, and V. Loreto. Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences USA*, 107(6):2403–2407, 2010.

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

[6] T. Belpaeme, J. Van Looveren, and L. Steels. The construction and acquisition of visual categories. In *European Workshop on Learning Robots*, pages 1–12. Springer, 1997.

[7] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[8] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.

[9] B. Berlin and E. A. Berlin. Aguaruna color categories 1. *American ethnologist*, 2(1):61–87, 1975.

[10] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1969.

[11] C. Bharathi and K. Pothi Reddy. Measuring critical flicker fusion frequency in human eye by utilizing sound card of the computer as dac. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 2015(3):48–50, 2015.

[12] C. P. Biggam. *Blue in Old English: An interdisciplinary semantic study*, volume 110. Rodopi, 1997.

[13] C. P. Biggam. *Grey in Old English: An interdisciplinary semantic study*. Runetree Press, 1998.

[14] C. P. Biggam. *The semantics of colour: A historical approach*. Cambridge University Press, 2012.

[15] D. Bimler. From color naming to a language space: An analysis of data from the world color survey. *Journal of Cognition and Culture*, 7(3):173–199, 2007.

[16] D. L. Bimler. Intimations of optimality: Extensions of simulation testing of color-language hypotheses. *Behavioral and Brain Sciences*, 28(4):489–490, 2005.

[17] D. M. Blei, M. I. Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[18] V. A. Bochko and K. A. Jameson. Investigating potential human tetrachromacy in individuals with tetrachromat genotypes using multispectral techniques. *Electronic Imaging*, 2018(14):1–12, 2018.

[19] J. Bowmaker and Y. Kunz. Ultraviolet receptors, tetrachromatic colour vision and retinal mosaics in the brown trout (salmo trutta): age-dependent changes. *Vision research*, 27(12):2101–2108, 1987.

[20] P. R. Boyce, N. H. Eklund, and S. N. Simpson. Individual lighting control: task performance, mood, and illuminance. *Journal of the Illuminating Engineering Society*, 29(1):131–142, 2000.

[21] D. Centola and A. Baronchelli. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences USA*, 112(7):1989–1994, 2015.

[22] S. Cho, A. Han, M. H. Taylor, A. C. Huck, A. M. Mishler, K. L. Mattal, C. A. Barker, and H.-S. Seo. Blue lighting decreases the amount of food consumed in men, but not in women. *Appetite*, 85:111–117, 2015.

[23] R. S. Cook, P. Kay, and T. Regier. The world color survey database. In *Handbook of categorization in cognitive science*, pages 223–241. Elsevier, 2005.

[24] P. J. Corr, A. D. Pickering, and J. A. Gray. Sociability/impulsivity and caffeine-induced arousal: Critical flicker/fusion frequency and procedural learning. *Personality and Individual Differences*, 18(6):713–730, 1995.

[25] B. De Vylder and K. Tuyls. How to reach linguistic consensus: A proof of convergence for the naming game. *Journal of theoretical biology*, 242(4):818–831, 2006.

[26] D. J. Dillon. Variability of flicker-fusion thresholds. *Perceptual and motor skills*, 17(3):666–666, 1963.

[27] I. Douven, S. Wenmackers, Y. Jraissati, and L. Decock. Measuring graded membership: the case of color. *Cognitive science*, 41(3):686–722, 2017.

[28] K. J. Emery, V. J. Volbrecht, D. H. Peterzell, and M. A. Webster. Variations in normal color vision. vi. factors underlying individual differences in hue scaling and their implications for models of color appearance. *Vision research*, 141:51–65, 2017.

[29] H. S. Fairman, M. H. Brill, and H. Hemmendinger. How the cie 1931 color-matching functions were derived from wright-guild data. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 22(1):11–23, 1997.

[30] N. Fider and N. L. Komarova. Quantitative study of color category boundaries. *JOSA A*, 35(4):B165–B183, 2018.

[31] N. Fider, L. Narens, K. A. Jameson, and N. L. Komarova. Quantitative approach for defining basic color terms and color category best exemplars. *JOSA A*, 34(8):1285–1300, 2017.

[32] N. A. Fider and N. L. Komarova. Differences in color categorization manifested by males and females: a quantitative world color survey study. *Palgrave Communications*, 5(1):1–10, 2019.

[33] L. Geiger. *Contributions to the History of the Development of the Human Race*, volume 12. Trübner & Company, 1880.

[34] S. J. Gershman and D. M. Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

[35] E. Gibson, R. Futrell, J. Jara-Ettinger, K. Mahowald, L. Bergen, S. Ratnasingam, M. Gibson, S. T. Piantadosi, and B. R. Conway. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences USA*, 114(40):10785–10790, 2017.

[36] W. E. Gladstone. *Studies on Homer and the Homeric age*, volume 1. Oxford University Press Oxford, 1858.

[37] H. Gleason. An introduction to descriptive linguistics (revised ed.)(new york). *Holt, Rinehart and*, 1961.

[38] M. Gooyabadi and K. Joe. Colorsims 2.0: An extension to the python package for evolving linguistic color naming conventions applied to a population of agents. Technical report, Technical Report Series# MBS18-02. Institute for Mathematical Behavioral Sciences, University of California, Irvine, Irvine, CA, USA, 2015.

[39] M. Gooyabadi, K. Joe, and L. Narens. Further evolution of natural categorization systems: an approach to evolving color concepts. *JOSA A*, 36(2):159–172, 2019.

[40] C. L. Hardin and L. Maffi. *Color categories in thought and language*. Cambridge University Press, 1997.

[41] F. I. Hárosi and Y. Hashimoto. Ultraviolet visual pigment in a vertebrate: a tetrachromatic cone system in the dace. *Science*, 222(4627):1021–1023, 1983.

[42] A. Hasenbeck, S. Cho, J.-F. Meullenet, T. Tokar, F. Yang, E. A. Huddleston, and H.-S. Seo. Color and illuminance level of lighting can modulate willingness to eat bell peppers. *Journal of the Science of Food and Agriculture*, 94(10):2049–2056, 2014.

[43] E. R. Heider. Universals in color naming and memory. *Journal of experimental psychology*, 93(1):10, 1972.

[44] E. R. Heider and D. C. Olivier. The structure of the color space in naming and memory for two languages. *Cognitive psychology*, 3(2):337–354, 1972.

[45] C. Hill. Named and unnamed spaces: color, kin, and the environment in umpila. *The Senses and Society*, 6(1):57–67, 2011.

[46] D. D. Hoffman and M. Singh. Computational evolutionary perception. *Perception*, 41(9):1073–1091, 2012.

[47] G. Hoffmann, V. Gufler, A. Griesmacher, C. Bartenbach, M. Canazei, S. Staggl, and W. Schobersberger. Effects of variable lighting intensities and colour temperatures on sulphatoxymelatonin and subjective mood in an experimental office workplace. *Applied Ergonomics*, 39(6):719–728, 2008.

[48] M. C. Hughes and E. Sudderth. Memoized online variational inference for dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 1133–1141, 2013.

[49] L. M. Hurvich and D. Jameson. An opponent-process theory of color vision. *Psychological review*, 64(6p1):384, 1957.

[50] K. Ikeuchi. Lambertian reflectance. *Encyclopedia of Computer Vision*, pages 441–443, 2014.

[51] K. Jameson and R. G. D'Andrade. 14 it's not really red, green, yellow, blue: an inquiry into perceptual color space. *Color categories in thought and language*, page 295, 1997.

[52] K. Jameson, N. Komarova, S. Tauber, and L. Narens. New results on simulated color categorization behaviors using realistic perceptual models, heterogeneous observers and pragmatic communication constraints., 2011. In *Presentation at The 17th annual meeting of the Cognitive Science Association for Interdisciplinary Learning*, 2011.

[53] K. A. Jameson. What saunders and van brakel chose to ignore in color and cognition research. *Behavioral and Brain Sciences*, 20(2):195–196, 1997.

[54] K. A. Jameson. On the role of culture in color naming: remarks on the articles of paramei, kay, roberson, and hardin on the topic of cognition, culture, and color experience. *Cross-Cultural Research*, 39(1):88–106, 2005.

[55] K. A. Jameson. Tetrachromatic color vision. In *The Oxford companion to consciousness*, pages 155–158. Oxford Press Oxford, 2009.

[56] K. A. Jameson and N. Alvarado. The relational correspondence between category exemplars and names. *Philosophical Psychology*, 16(1):25–49, 2003.

[57] K. A. Jameson, S. M. Highnote, and L. M. Wasserman. Richer color experience in observers with multiple photopigment opsin genes. *Psychonomic bulletin & review*, 8(2):244–261, 2001.

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

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

[60] K. A. Jameson and N. L. Komarova. Evolutionary models of color categorization based on realistic observer models and population heterogeneity. *Journal of Vision*, 10(15):26–26, 2010.

[61] K. A. Jameson, A. D. Winkler, and K. Goldfarb. Art, interpersonal comparisons of color experience, and potential tetrachromacy. *Electronic Imaging*, 2016(16):1–12, 2016.

[62] K. A. Jameson, A. D. Winkler, C. Herrera, and K. Goldfarb. The veridicality of color: A case study of potential human tetrachromacy. *Technical Report Series MBS 14-02. Institute for Mathematical Behavioral Sciences University of California at Irvine. Irvine, CA, USA*, 2014.

[63] I. Juricevic and M. A. Webster. Variations in normal color vision. v. simulations of adaptation to natural color environments. *Visual neuroscience*, 26(1):133–145, 2009.

[64] L. Kaufman and P. J. Rousseeuw. Finding groups in data: An introduction to cluster analysis–john wiley & sons. *Inc., New York*, 1990.

[65] P. Kay, B. Berlin, L. Maffi, W. Merrifield, et al. Color naming across languages. *Color categories in thought and language*, 21:58, 1997.

[66] P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, and R. Cook. *The World Color Survey*. CSLI Publications Stanford, 2009.

[67] P. Kay, B. Berlin, and W. Merrifield. Biocultural implications of systems of color naming. *Journal of Linguistic Anthropology*, 1(1):12–25, 1991.

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

[69] P. Kay and L. Maffi. Color appearance and the emergence and evolution of basic color lexicons. *American anthropologist*, 101(4):743–760, 1999.

[70] P. Kay and C. K. McDaniel. The linguistic significance of the meanings of basic color terms. *Language*, pages 610–646, 1978.

[71] P. Kay and T. Regier. Language, thought and color: recent developments. *Trends in cognitive sciences*, 10(2):51–54, 2006.

[72] N. L. Komarova and K. A. Jameson. Population heterogeneity and color stimulus heterogeneity in agent-based color categorization. *Journal of theoretical Biology*, 253(4):680–700, 2008.

[73] N. L. Komarova, K. A. Jameson, and L. Narens. Evolutionary models of color categorization based on discrimination. *Journal of Mathematical Psychology*, 51(6):359–382, 2007.

[74] I. Kuriki, R. Lange, Y. Muto, A. M. Brown, K. Fukuda, R. Tokunaga, D. T. Lindsey, K. Uchikawa, and S. Shioiri. The modern japanese color lexicon. *Journal of vision*, 17(3):1–1, 2017.

[75] M. R. Leek. Adaptive procedures in psychophysical research. *Perception & psychophysics*, 63(8):1279–1292, 2001.

[76] S. C. Levinson. Yélî dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, 10(1):3–55, 2000.

[77] D. T. Lindsey and A. M. Brown. Universality of color names. *Proceedings of the National Academy of Sciences*, 103(44):16608–16613, 2006.

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

[79] D. T. Lindsey, A. M. Brown, D. H. Brainard, and C. L. Apicella. Hunter-gatherer color naming provides new insight into the evolution of color terms. *Current Biology*, 25(18):2441–2446, 2015.

[80] V. Loreto, A. Mukherjee, and F. Tria. On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, 109(18):6819–6824, 2012.

[81] V. Loreto and L. Steels. Social dynamics: Emergence of language. *Nature Physics*, 3(11):758, 2007.

[82] J. A. Lucy. 15 the linguistics of" color. *Color categories in thought and language*, page 320, 1997.

[83] J. A. Lucy and R. A. Shweder. The effect of incidental conversation on memory for focal colors. *American Anthropologist*, 90(4):923–931, 1988.

[84] J. Luttinen. Bayespy: variational bayesian inference in python. *The Journal of Machine Learning Research*, 17(1):1419–1424, 2016.

[85] J. Lyons. Colour in language. *Colour: Art and science*, 194223, 1995.

[86] R. E. MacLaury. *Color and cognition in Mesoamerica: Constructing categories as vantages*. University of Texas press, 1997.

[87] G. Malkoc, P. Kay, and M. A. Webster. Variations in normal color vision. iv. binary hues and hue scaling. *JOSA A*, 22(10):2154–2168, 2005.

[88] L. T. Maloney. Physics-based approaches to modeling surface color perception. *Color vision: From genes to perception*, pages 387–416, 1999.

[89] M. G. Masi, L. Peretto, R. Tinarelli, and L. Rovati. A pupil size measurement system for the analysis of the impact of flicker on human being. In *Proc. of the 16th IMEKO TC-4 Symposium*, pages 419–424, 2008.

[90] S. L. Merbs and J. Nathans. Absorption spectra of the hybrid pigments responsible for anomalous color vision. *Science*, 258(5081):464–466, 1992.

[91] J. Mollon and B. Regan. Cambridge colour test handbook. *Cambridge: Cambridge Research Systems*, 2000.

[92] D. Mylonas and L. MacDonald. Augmenting basic colour terms in english. *Color Research & Application*, 41(1):32–42, 2016.

[93] L. Narens, K. A. Jameson, N. L. Komarova, and S. Tauber. Language, categorization, and convention. *Advances in Complex Systems*, 15(03n04):1150022, 2012.

[94] K. Nassau. *The physics and chemistry of color: the fifteen causes of color*. 2001.

[95] J. Nathans, T. P. Piantanida, R. L. Eddy, T. B. Shows, and D. S. Hogness. Molecular genetics of inherited variation in human color vision. *Science*, 232(4747):203–210, 1986.

[96] M. Ni, E. B. Sudderth, and M. Hughes. Variational inference for beta-bernoulli dirichlet process mixture models. 2015.

[97] D. Oberfeld, H. Hecht, U. Allendorf, and F. Wickelmaier. Ambient lighting modifies the flavor of wine. *Journal of Sensory Studies*, 24(6):797–832, 2009.

[98] P. Orbanz and Y. W. Teh. Bayesian nonparametric models. *Encyclopedia of machine learning*, (1), 2010.

[99] G. V. Paramei. Singing the russian blues: An argument for culturally basic color terms. *Cross-cultural research*, 39(1):10–38, 2005.

[100] G. V. Paramei. Russian 'blues'. *Anthropology of color: Interdisciplinary multilevel modeling*, 75, 2007.

[101] G. V. Paramei. Color discrimination across four life decades assessed by the cambridge colour test. *JOSA A*, 29(2):A290–A297, 2012.

[102] J. Park, S. Tauber, K. A. Jameson, and L. Narens. The evolution of shared concepts in changing populations. *Review of Philosophy and Psychology*, 10(3):479–498, 2019.

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

[104] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[105] V. F. Ray. Techniques and problems in the study of human color perception. *Southwestern Journal of Anthropology*, 8(3):251–259, 1952.

[106] V. F. Ray. Human color perception and behavioral response. *Transactions of the New York Academy of Sciences*, 16(2):98–105, 1953.

[107] B. C. Regan, J. Reffin, and J. D. Mollon. Luminance noise and the rapid determination of discrimination ellipses in colour deficiency. *Vision research*, 34(10):1279–1299, 1994.

[108] T. Regier, P. Kay, and R. S. Cook. Focal colors are universal after all. *Proceedings of the National Academy of Sciences USA*, 102(23):8386–8391, 2005.

[109] T. Regier, P. Kay, and R. S. Cook. Universal foci and varying boundaries in linguistic color categories. In *Proceedings of the 27th Meeting of the Cognitive Science Society*, pages 1827–1832. Citeseer, 2005.

[110] T. Regier, P. Kay, A. L. Gilbert, and R. B. Ivry. Which side are you on, anyway? *Words and the mind: How words capture human experience*, pages 165–182, 2010.

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

[112] D. Roberson. Color categories are culturally diverse in cognition as well as in language. *Cross-Cultural Research*, 39(1):56–71, 2005.

[113] D. Roberson, J. Davidoff, I. R. Davies, and L. R. Shapiro. Color categories: Evidence for the cultural relativity hypothesis. *Cognitive psychology*, 50(4):378–411, 2005.

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

[115] M. Rodriguez-Carmona. Genetics of photoreceptors, genetics and color vision deficiencies, genes of cone photopigments, genes and cones. *Encyclopedia of Color Science and Technology*, pages 1–6, 2014.

[116] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[117] B. A. Saunders and J. Van Brakel. Are there nontrivial constraints on colour categorization? *Behavioral and brain sciences*, 20(2):167–179, 1997.

[118] A. R. Seitz, J. E. Nanez Sr, S. R. Holloway, and T. Watanabe. Perceptual learning of motion leads to faster flicker perception. *PLoS One*, 1(1):e28, 2006.

[119] L. T. Sharpe, A. Stockman, H. Jägle, and J. Nathans. Opsin genes, cone photopigments, color vision, and color blindness. *Color vision: From genes to perception*, 351, 1999.

[120] B. R. Shelton, M. C. Picardi, and D. M. Green. Comparison of three adaptive psychophysical procedures. *The Journal of the Acoustical Society of America*, 71(6):1527–1533, 1982.

[121] B. Skryms. *Evolution of the social contract*. Cambridge University Press, 1996.

[122] K. Smith, S. Kirby, and H. Brighton. Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–386, 2003.

[123] J. Stanlaw. 11 two observations on culture contact and the japanese color nomenclature system. *Color categories in thought and language*, page 240, 1997.

[124] J. Stanlaw. Language, contact, and vantages: fifteen hundred years of japanese color terms. *Language sciences*, 32(2):196–224, 2010.

[125] L. Steels. Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7):308–312, 2003.

[126] L. Steels. Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4):339–356, 2011.

[127] L. Steels, T. Belpaeme, et al. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, 28(4):469–488, 2005.

[128] R. Steingrimsson. 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, 2012.

[129] W. S. Stiles and J. M. Burch. Npl colour-matching investigation: final report (1958). *Optica Acta: International Journal of Optics*, 6(1):1–26, 1959.

[130] J. Stillman. A comparison of three adaptive psychophysical procedures using inexperienced listeners. *Perception & psychophysics*, 46(4):345–350, 1989.

[131] S. Tauber. Colorsims: A python package for evolving linguistic color naming conventions within a population of simulated agents. Technical report, Technical Report Series# MBS15-01. Institute for Mathematical Behavioral Sciences, University of California, Irvine, Irvine, CA, USA, 2015.

[132] F. Tria, A. Mukherjee, A. Baronchelli, A. Puglisi, and V. Loreto. A fast no-rejection algorithm for the category game. *Journal of Computational Science*, 2(4):316–323, 2011.

[133] K. Uchikawa and R. M. Boynton. Categorical color perception of japanese observers: Comparison with that of americans. *Vision Research*, 27(10):1825–1833, 1987.

[134] J. A. Veitch and G. R. Newsham. Lighting quality and energy-efficiency effects on task performance, mood, health, satisfaction, and comfort. *Journal of the Illuminating Engineering Society*, 27(1):107–129, 1998.

[135] M. Vianya-Estopa, W. A. Douthwaite, K. Pesudovs, B. A. Noble, and D. B. Elliott. Development of a critical flicker/fusion frequency test for potential vision testing in media opacities. *Optometry and vision science*, 81(12):905–910, 2004.

[136] M. A. Webster, I. Juricevic, and K. C. McDermott. Simulations of adaptation and color appearance in observers with varying spectral sensitivity. *Ophthalmic and Physiological Optics*, 30(5):602–610, 2010.

[137] M. A. Webster and P. Kay. Individual and population differences in focal colors. *The anthropology of color*, 2007.

[138] M. A. Webster, E. Miyahara, G. Malkoc, and V. E. Raker. Variations in normal color vision. i. cone-opponent axes. *JOSA A*, 17(9):1535–1544, 2000.

[139] M. A. Webster, E. Miyahara, G. Malkoc, and V. E. Raker. Variations in normal color vision. ii. unique hues. *JOSA A*, 17(9):1545–1555, 2000.

[140] M. A. Webster, S. M. Webster, S. Bharadwaj, R. Verma, J. Jaikumar, G. Madan, and E. Vaithilingham. Variations in normal color vision. iii. unique hues in indian and united states observers. *JOSA A*, 19(10):1951–1962, 2002.

[141] N. Zaslavsky, C. Kemp, T. Regier, and N. Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.

Appendix A

Pairwise Agent Comparison Measure

The pairwise agent comparison measure (C_r) is defined to be an alternative method for evaluating the level of agreement of a population's color categorization at a given time. Where as the global agreement measure (A_r) aggregates agreement across the set of color chips (see Equation 2.1 in Section 2.2.3), pairwise agent comparison aggregates agreement across all unique pairs of agents. C_r is formally defined as follows:

$$C_r = \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k=1}^{320} [W_{ik} = W_{jk}]}{\frac{(N-1)N}{2}} \tag{A.1}$$

where i and j are different agents from the population, N is the population size, k is the chip number, r is the number of rounds, W_{ik} is the word that agent i assigned to color chip k, the number of possible pairs of agents is given by $\frac{(N-1)N}{2}$, and $[W_{ik} = W_{jk}] = \begin{cases} 1 & \text{if } W_{ik} = W_{jk} \\ 0 & \text{if } W_{ik} \neq W_{jk} \end{cases}$.

Appendix B

Schematic Similarity: Proof of Metric

In mathematics, a *metric* is defined as a function that describes the distance between pairs of elements within a set. A metric $d(x, y)$ defined on the set S also satisfies the following axioms, such that $\forall x, y, z \in S$:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \iff x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

Our measure, called *Schematic Similarity* (SS; see Section 4.2.4.3), describes an inverse distance relationship between pairs of elements (i.e. WCS participants). Therefore, we let $d(x, y) = 1 - SS(x, y)$ be our distance function. Here we provide a formal proof, organized according to the four axioms listed above, that $d(x, y)$ is a mathematical metric.

Theorem B.1. *The distance function $d(x, y) = 1 - SS(x, y)$ defined on the set X, which is the set of all WCS participants, is a metric.*

Proof. Let $x, y, z \in X$. WLOG, assume that x is the *Reference Participant* (i.e. participant using fewer color terms) and y is the *Other Participant*.

Axiom 1: d(x,y) $\geq$ 0

Let M be the set of mapped terms between participants x and y. (See Section 4.2.4.3 for more information on how to generate this set.)

By definition, $d(x, y) = 1 - SS(x, y)$.

Since $0 \leq SS(x, y) \leq 1 \; \forall x, y \in X$, then $d(x, y) \geq 0 \; \forall x, y \in X$.

Axiom 2: d(x,y)=0 $\iff$ x=y

($\Longleftarrow$) Assume $x = y$ (i.e. x and y are the same participant).

Then the set of mapped terms

$$M = \bigcup_{i=1}^{n} \{(i, i)\}$$

where n = number of terms x used and each i represents an individual term.

Since $x = y$, $\forall (i, j) \in M$, $i = j$.

Therefore, $|T_i^x \cap T_i^x| = |T_i^x \cup T_i^x|$ and $e_{T_i^x} = 0$ because $U = \emptyset$. Hence,

$$\begin{aligned}
d(x, x) &= 1 - SS(x, x) \\
&= 1 - \sum_{(i,i) \in M} \frac{1}{n} \frac{|T_i^x \cap T_i^x|}{|T_i^x \cup T_i^x| + e_{T_i^x}} \\
&= 1 - \sum_{(i,i) \in M} \frac{1}{n} \frac{|T_i^x \cap T_i^x|}{|T_i^x \cup T_i^x|} \\
&= 1 - \sum_{(i,i) \in M} \frac{1}{n} * 1 \\
&= 1 - \frac{n}{n} \\
&= 1 - 1 \\
&= 0
\end{aligned}$$

($\implies$) Assume that $d(x, y) = 0$. Then,

$$\begin{aligned} 1 - SS(x, y) &= 0 \\ SS(x, y) &= 1 \\ \sum_{(i,j)\in M} \frac{1}{n} \frac{|T_i^x \cap T_j^y|}{|T_i^x \cup T_j^y| + e_{T_i^x}} &= 1 \\ \sum_{(i,j)\in M} \frac{|T_i^x \cap T_j^y|}{|T_i^x \cup T_j^y| + e_{T_i^x}} &= n \end{aligned}$$

Since $0 \leq SS(x, y) \leq 1$ and $|M| = n$, then

$$\frac{|T_i^x \cap T_j^y|}{|T_i^x \cup T_j^y| + e_{T_i^x}} = 1, \quad \forall (i, j) \in M$$

Therefore, $|T_i^x \cap T_j^y| = |T_i^x \cup T_j^y| + e_{T_i^x}$.

By construction, $e_{T_i^x} \in \mathbb{Z}^+ \cup \{0\}$.

Suppose $e_{T_i^x} > 0$.

Then $|T_i^x \cap T_j^y| > |T_i^x \cup T_j^y|$, which is a contradiction.

Therefore, $e_{T_i^x} = 0$.

It then follows that $|T_i^x \cap T_j^y| = |T_i^x \cup T_j^y|$. In other words, for every term i used by participant x and its mapped term j used by participant y, the set of chips being called i and j by x and y, respectively, are exactly equal. That is, terms i and j refer to the exact same regions of the color space.

Therefore, the mapping from terms used by x and terms used by y is a bijection.

Since x and y use the same number of color terms and each mapped pair of terms refers to the exact same region of the color space, participants x and y have identical partitions of the color space. Therefore, $x = y$.

Axiom 3: d(x,y) = d(y,x)

Recall that x is assumed to be the *Reference Participant (RP)* and y is assumed to be the *Other Participant* (OP).

Then $SS(x, y) = SS(y, x)$ because *Schematic Similarity* only cares about who is the RP and who is the OP, not about the order in which they are given to the function. Therefore,

$$1 - d(x, y) = SS(x, y) = SS(y, x) = 1 - d(y, x)$$

$$1 - d(x, y) = 1 - d(y, x)$$

$$d(x, y) = d(y, x)$$

Axiom 4: d(x,z) ≤ d(x,y) + d(y,z)

Due to the formulation of *Schematic Similarity*, it is impossible to prove the Triangle Inequality of our distance metric analytically. Therefore, we computed $d(x, y)$, $d(y, z)$, and $d(x, z)$ $\forall (x, y, z) \in X \times X \times X$ and checked $d(x, z) \leq d(x, y) + d(y, z)$. We found this inequality to hold for every combination of WCS participants. For our purposes, this is sufficient evidence that the Triangle Inequality holds for our distance metric d. □

Appendix C

Tables of Results from Experimental Flicker Study

Included in this appendix are tables which contain further information regarding results that were reported in the Results section of Chapter 5 (Section 5.3).

Tables C.1 and C.2 are both presented in support of the finding that the use of a truncated/LED illuminant worsens the metamerism between the two RHEM stimuli. Table C.1 relays the results of a two-sample hypothesis test which compares the average minimzied flicker setting for the RHEM A and B under the D5000 pair versus the RHEM A and B under a truncated illuminant pair. Every hypothesis test concluded that the minimized flicker settings for the two pairs were significantly different. In fact, subjects always set their minimized flicker settings for pairs using a truncated illuminant higher than for the pair using the D5000. The interpretation of this result in context of the experiment is that the perceived difference between RHEM A and B is more pronounced when illuminanted by an LED than by the standard. For brevity, spectral pairs were given a shortened name in the table. The shortened names can be translated as follows:

- D5000AB = D5000xA and D5000xB
- DTAB = DTxA and DTxB
- DT*AB = DT*xA and DT*xB

Table C.2 contains ΔE values for the different spectral pairs listed in the table, both in CIE LAB and LUV color spaces. Only the truncated illuminants with X deviation values 100%, 102%, and 103% are displayed in the table because those were the only truncated illuminants that were a best match to at least one subject in Stimulus Condition 1. The values in this table demonstrate how the distance between the RHEM A and B spectra becomes greater when convolved with a truncated illuminant as opposed to the D5000. This supports the results of the hypothesis tests in Table C.1 which show that subjects are perceiving the indicator pair as more different when under an LED light.

Tables C.3 and C.4 are both presented in support of the finding that spectral pairs which represent either RHEM A or B being convolved with the D5000 and one of the truncated illuminants all have minimized flicker settings which are significantly greater than 1 frm/sec. This means that for all of these spectral pairs, subjects could perceive a difference between the two stimuli. Therefore, from these results we can conclude that changing the illuminant from the D5000 to one of its LED approximates is affecting the color that subjects perceive RHEM A or B as. The ΔE values reported in Table C.4 also support this finding.

Subject ID	Spectral Pairs	Average MFSs	P-value	Different?
01	D5000AB // DTAB	32.4 // 40.8	2.24×10^{-12}	Yes
	D5000AB // DT*AB	32.4 // 40.6	9.27×10^{-9}	Yes
02	D5000AB // DTAB	21.5 // 31.5	2.86×10^{-11}	Yes
	D5000AB // DT*AB	21.5 // 33.3	3.12×10^{-12}	Yes
03	D5000AB // DTAB	36.4 // 44.55	0.005138	Yes
04	D5000AB // DTAB	28.5 // 39.1	1.9×10^{-7}	Yes
09	D5000AB // DTAB	23.8 // 34.4	9.69×10^{-12}	Yes
11	D5000AB // DTAB	22.1 // 32.3	3.42×10^{-7}	Yes
	D5000AB // DT*AB	22.1 // 32.7	1.06×10^{-8}	Yes

Table C.1: Results from the two-sample hypothesis test $H_0 : \mu_{MFS\_1} = \mu_{MFS\_2}$, $H_1 : \mu_{MFS\_1} > \mu_{MFS\_2}$ for the 6 subjects who participated in Stimulus Condition 2. Based on the results from the hypothesis test, we can infer whether the change is illuminant (from the standard D5000 to an LED approximate) is affecting the strength of the metamerism between RHEM A and B. The last column of the table details whether the difference between RHEM A and B is being dramatized by the truncated illumiant.

Spectral Pair	ΔE^*_{ab}	ΔE^*_{uv}
D5000xA // D5000xB	0.7969	2.6272
DTxA // DTxB	6.3919	9.4115
DT_{102}xA // DT_{102}xB	6.8816	9.4115
DT_{103}xA // DT_{103}xB	5.5910	12.2568

Table C.2: ΔE values for spectral pairs representing either RHEM A or B being illuminated by the same illuminant. ΔE was computed for each spectral pair both in CIE LAB and LUV color spaces.

Subject ID	Spectral Pair	Average MFS	P-value	Metameric?
01	D5000xA // DTxA	34.1	4.62×10^{-14}	No
	D5000xA // DT*xA	31.4	5.34×10^{-13}	No
	D5000xB // DTxB	21.7	3.61×10^{-11}	No
	D5000xB // DT*xB	27.5	3.43×10^{-11}	No
02	D5000xA // DTxA	28.5	6.74×10^{-11}	No
	D5000xA // DT*xA	25.4	3.84×10^{-12}	No
	D5000xB // DTxB	14.7	6.60×10^{-11}	No
	D5000xB // DT*xB	16.9	1.80×10^{-12}	No
03	D5000xA // DTxA	39.2	0	No
	D5000xB // DTxB	24	0	No
04	D5000xA // DTxA	33.05	0	No
	D5000xB // DTxB	17.6	9.23×10^{-8}	No
09	D5000xA // DTxA	26.3	4.44×10^{-14}	No
	D5000xB // DTxB	19.2	2.69×10^{-14}	No
11	D5000xA // DTxA	25.2	3.67×10^{-12}	No
	D5000xA // DT*xA	24.2	3.03×10^{-11}	No
	D5000xB // DTxB	15.1	7.31×10^{-10}	No
	D5000xB // DT*xB	17.1	8.95×10^{-10}	No

Table C.3: Results from the hypothesis test $H_0 : \mu_{MFS} = 1, H_1 : \mu_{MFS} > 1$ for the 6 subjects who participated in Stimulus Condition 2 for a subset of spectral pairs. Based on the results from the hypothesis test, we can infer whether the change in illuminant (from standard D5000 to an LED approximate) is affecting the way the subjects are perceiving the RHEM A and RHEM B spectra. The last column of the table details whether the participant perceived the two spectra as metamers or not.

Spectral Pair	ΔE^*_{ab}	ΔE^*_{uv}
D5000xA // DTxA	4.5127	13.2018
D5000xA // DT_{102}xA	4.0605	11.4929
D5000xA // DT_{103}xA	3.1651	11.1798
D5000xB // DTxB	1.1468	9.8273
D5000xB // DT_{102}xB	2.0302	6.3949
D5000xB // DT_{103}xB	1.6489	3.7094

Table C.4: ΔE values for spectral pairs representing RHEM A under different illuminants or RHEM B under different illuminants. ΔE was computed for each spectral pair both in CIE LAB and LUV color spaces.

Appendix D

Calculating the Tristimulus Values and LAB & LUV Coordinates

All stimuli used in the study described in Chapter 5 were first measured using a *Mathematica* program written by Dr. Tim Satalich which employed functions from the Ocean Optics software development kit. Stimuli were displayed through the OL-490 into an integrating sphere. A fiber optic cable was used to connect the Ocean Optics Flame Spectrometer to the north port of the integrating sphere. (Figure 5.5a depicts the measurement environment described above.)

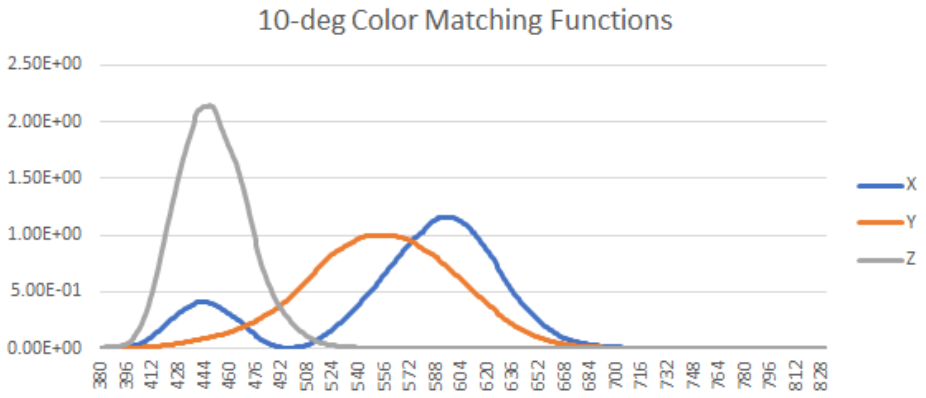

Figure D.1: Color matching functions of the CIE 10° standard observer model.

The resulting reflectance spectra were configured into an Nx401 matrix, where N is the number of stimuli, and 401 is the number of wavelengths (380–780 nm). This matrix was multiplied by a 401x3 matrix representing the L, M, and S chromatic response functions of the 1964 CIE 10° standard observer (see Figure D.1). The resulting Nx3 matrix provided the X, Y, and Z values of each stimulus.

The color matching functions used to compute the tristimulus values from reflectance spectra were taken from the Color and Vision Research Laboratory website (www.cvrl.org). The functions were the "10-deg XYZ CMFs transformed from the CIE (2006) 2-deg LMS cone fundamentals) with a 1 nm stepsize."

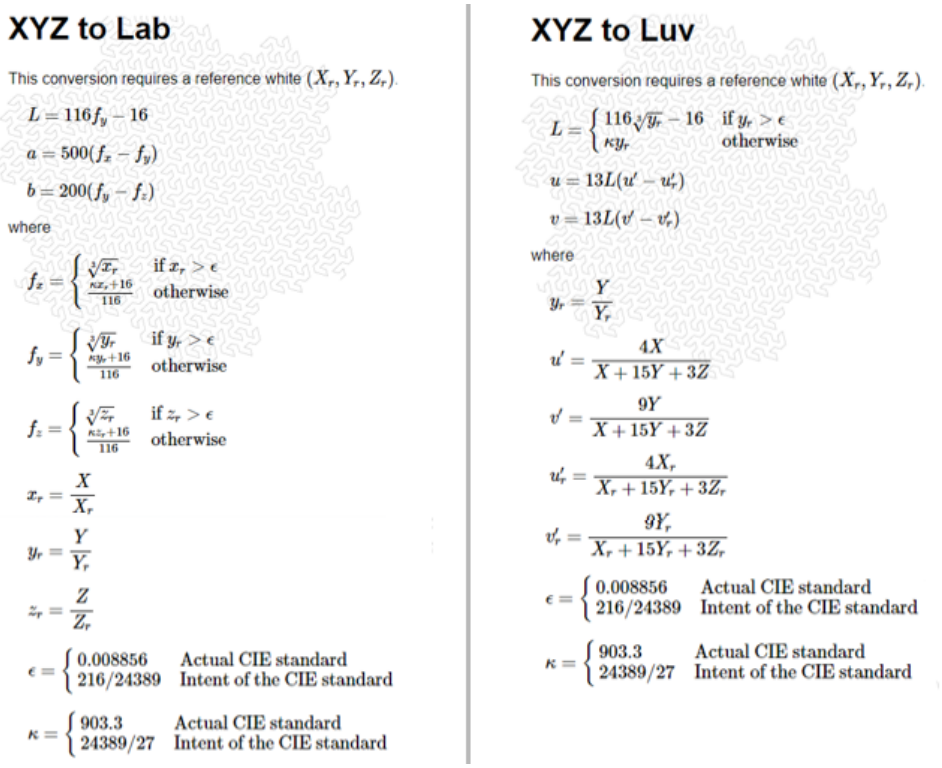

Figure D.2: Series of functions used to convert the tristimulus X, Y, Z values into CIE LAB and LUV coordinates. These functions are provided by brucelindbloom.com.

After using the above procedure to calculate the tristimulus values of the specta we measured with the Ocean Optics Flame spectrometer, we converted the X, Y, Z values to CIE LAB and LUV coordinates. We chose to convert the stimuli into LAB and LUV coordinates because

of their respective color space's property as a perceptual color space. In a perceptual color space, distances in the space map to similar distances in our perception, so a distance metric like ΔE can provide us with an estimate of how perceptually "far" two colors are from each other. As designed by the CIE, colors can be translated from one color space to another by a simple algebraic transformation (see Figure D.2). The functions used to transform the tristiumulus X, Y, Z values into the LAB and LUV coordinates were taken from Bruce Lindblom's website, www.brucelindbloom.com.