MEMORY LANE: EVALUATING FACTORS THAT CONTRIBUTE TO LONG-TERM EPISODIC MEMORY

By

KIMELE PERSAUD

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Psychology

Written under the direction of

Pernille Hemmer

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2018

ABSTRACT OF THE DISSERTATION

Memory Lane: Evaluating Factors that Contribute to Long-term Episodic Memory

by KIMELE PERSAUD

Dissertation Director:

Pernille Hemmer

Visual working (WM) and long-term memory (LTM) are intricately intertwined. As such, current theories and models of VWM have been extended to characterize behavior in long-term memory. For example, a popular framework for investigating VWM is the remember-guess paradigm, which suggests that information is either recalled with some noise, or is no longer retrievable and individuals resort to random guessing (e.g. Brady et al., 2013). This framework has been extended to include an additional factor that contributes to memory, namely interference from non-target information (a.k.a. misassociations; Lew et al, 2015). In this way, individuals recall information with noise, missassociate memories to other task relevant information, or guess randomly. The compilation of these studies has identified the contribution of memory fidelity, misassociations, and random guesses to recall performance.

Notably, the remember-guess framework stands in stark contrast to theoretical Bayesian models of memory, which suggests that prior knowledge and expectations for the statistical regularities of the environment influences recall from long-term memory (Hemmer & Steyvers, 2009b). The influence of prior knowledge is most prevalent when the stimuli in the memory tasks mirror the regularities of the natural world.

In this dissertation, I seek to challenge current theories of memory regarding the contribution of fidelity, misassociations, and random guesses to LTM, by evaluating the simultaneous contribution of prior knowledge. The combination of results from these studies suggest that prior knowledge plays a crucial recall in reconstruction from long-term episodic memory, and when prior knowledge is brought to the task of remembering, it alters the contribution of misassociations and random guessing to recall performance.

Acknowledgements

I would like express my deepest appreciation to my advisor, Dr. Pernille Hemmer, for her invaluable support and guidance throughout my graduate career. I owe a great deal of my success to her steadfast diligence and constant encouragement. I would also like to thank my committee members, Dr. Eileen Kowler, Dr. Jacob Feldman, and Dr. Ed Vul for their insightful comments and feedback which have culminated into making this dissertation as strong as it can be. Special thanks to the current and former members of the Priors and Memory (Prime) lab, without whom this work would not be possible. Specifically, I wish to thank Talia Robbins, Daniel Wall, Daljit Ahluwalia, Kevin Pei, Kierra Pean, Chrystal Spencer, and Aadarsh Kandevel. Also, I sincerely thank Dr. Matthew Stone, Brian McMahon, and Malihe Ahlikhani.

Importantly, I would like to thank God and my family (especially my mom, sister, and godmother) for walking with me through this insane, and exciting, and stressful, and fulfilling journey known as graduate school. I deeply thank my husband, Davon, for being my rock and encouraging me to keep moving forward in my academic career. I love you more than words could ever say.

I would be remiss if I did not also thank the hardworking staff in the psychology department and the Center for Cognitive Science here are Rutgers for their help over the years, namely Anne Sokolowski, Jo'Ann Meli, John Dixon, and Tamela Wilcox.

Chapter 2, in full, is a reprint of the material as it appears in: Persaud, K., & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. In P. Bello, M.Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the Cognitive Science Society (pp. 1162–1167).* Quebec City, CA:

Cognitive Science Society. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in partial, is a reprint of the material as it appears in: Hemmer, P., Persaud, K., McGovern, C., & Piantidosi, S. (2015). Shifting priors: Evaluating the cross-cultural influence of color expectations on episodic memory. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society. The dissertation author was the primary investigator and author of this work.

Chapter 4, in full is a reprint of the material as it appears in: Persaud, K. & Hemmer, P. (2016). The dynamics of fidelity over the time course of long-term memory. *Cognitive Psychology, 88,* 1-21. The dissertation author was the primary investigator and author of this paper.

Completion of this work was supported by the National Science Foundation Graduate Research Fellowship (NSF DGE 0937373), National Science Foundation Integrative Graduate Education and Research Traineeship (IGERT DGE 0549115), and the National Science Foundation CAREER Grant (1453276).

Dedication

This dissertation is dedicated to my husband and my children; my daughter Shayla and my unborn child (see you in August!). They are my motivation and constant reminder of why I must succeed. I want them to know that the sky is the limit and nothing is impossible to them that believe!

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

A particularly important question for memory research regards the nature of episodic memory over time is: what happens to memory traces as they transition from visual short-term/working memory into long-term memory and what factors contribute to long-term memory performance? These questions have significant implications for how long-term memory is theorized, and in turn operationalized in models of long-term memory. Various paradigms and accompanying models have been implemented to explain long-term memory, with some being derived from studies of visual working memory, under the assumption that processes and mechanisms of short-term and working memory also exist in long-term memory (Brady, Konkle, Gill, Oliva, & Alvarez, 2013; Donkin, Nosofsky, Gold, & Shiffrin, 2014; Huttenlocher, Hedges, & Vevea, 2000; Hemmer & Steyvers, 2009b; Persaud & Hemmer, 2016; Lew, Pashler, & Vul, 2015).

These studies have identified four factors that contribute to the reconstruction of information from long-term episodic memory, namely: prior knowledge, memory fidelity, random guessing and interference (a.k.a. misassociations) which results when non-target information stored in memory *interferes* with the retrieval of target information. Prior knowledge and expectations for the statistical regularities of the environment have been shown, on average, to improve recall from long-term memory (Hemmer & Steyvers, 2009a; 2009b). Random guessing contributes to recall performance when information reaches a low state of fidelity and is no longer retrievable from memory (Brady et al., 2013). Alternatively, when information is difficult to retrieve,

individuals might use other task relevant information (i.e., missassociate), before resorting to random guessing (Lew et al., 2015).

In this way, the role of each of these factors in long-term episodic memory has been studied relatively independent of one another (with the exception of random guessing) in terms of their impact on memory. For example, Lew and colleagues (2015) evaluate the role of misassociations and random guessing in long-term memory, but not the influence of prior knowledge. Similarly, Brady and colleagues (2013) evaluate the role of fidelity and random guessing in long-term memory, but not the influence of interference in the form of misassociations. However, for certain stimulus environment, particularly when the environment reflects features of the real world, the contribution of these factors may be intricately intertwined.

Therefore, the work presented in this dissertation seeks to address the question of what happens to information over time, while simultaneously evaluating the combined contribution of these four factors to long-term memory performance. In what follows is a brief overview of each topic that will be discussed and the corresponding chapters in which they can be found. In the chapters 2-4, I will present published research from three studies.

Chapter 2 details a study that empirically and computationally assessed the role of prior knowledge in long-term episodic memory for color and appears in *Proceedings of the Annual Meeting of the Cognitive Science Society.* This work was presented at the Cognitive Science Society Conference and received the Glushko Student Travel Award. This work demonstrated that people's categorical knowledge and expectations influence episodic memory, and that this reconstructive process can be simulated with a generative

Bayesian model. Chapter 3 discusses a provisional cross-cultural study that extended the findings from chapter 2, and demonstrated that the use of prior knowledge may be a general mechanism of episodic memory. This work also appeared in the *Proceedings of the Annual Meeting of the Cognitive Science Society* and partially in *i-Perception.* The combination of the studies in Chapters 2 and 3 illustrate that the role of prior knowledge should not be ignored in theories and models of long-term episodic memory.

The research presented in Chapter 4 explored the role of memory fidelity, prior knowledge, and random guessing in long-term memory and compared the performance of current models of memory. This work appears in *Cognitive Psychology* and was presented at the *Annual Meeting of the Mathematical Society*. The results from this work suggested that there are factors that influence memory such as prior knowledge and other factors that result in low-state fidelity that have been ignored in previous memory models. In previous models, the influence of these factors has erroneously been attributed to random guessing. Also, certain analytical practices (e.g. evaluating aggregated error distributions) used in past models obscured important contributions of factors to memory, such as prior knowledge. This work made transformative discoveries to how memory works and exposed a major flaw in current practices for evaluating memory data.

Lastly, Chapter 5 presents new work evaluating the contribution of prior knowledge, interference in the form of misassociations, and random guessing in long-term memory. The results from this work demonstrated that a large portion of errors in memory for meaningful stimuli, resulted from misassociations and prior knowledge, *not* random guessing. These results supported the hypothesis that there is little to no random guessing in LTM for semantically associated, ecologically valid stimuli. The combination of all

studies discussed in this dissertation provides a comprehensive understanding of long-term memory and the factors that contribute to memory performance.

Chapter 2: Prior knowledge and Memory

The Influence of Knowledge and Expectations for Color on Episodic Memory

Kimele Persaud and Pernille Hemmer (2014). *Proceedings of the 36th Annual Conference of the Cognitive Science Society*

K. Persaud and the advisor, P. Hemmer, developed the study concept and study design together. Stimulus creation, testing and data collection were performed by K. Persaud. K. Persaud performed the data analysis and interpretation, which were then reviewed by the advisor P. Hemmer. K. Persaud and P. Hemmer developed and implemented the model together. K. Persaud drafted the manuscript. After the manuscript was drafted, K. Persaud and the advisor, P. Hemmer, revised the manuscript. K. Persaud implemented all critical revisions in response to reviewer comments.

Abstract

Expectations learned from our environment are known to exert strong influences on episodic memory. Furthermore, people have prior expectations for universal color labels and their associated hue space—a salient property of the environment. In three experiments, we assessed peoples' color naming preferences, and expectation for color. Using a novel experimental paradigm, we then assessed free recall for color. We found that people's color naming preferences were consistent with the universal color terms (Berlin & Kay, 1969), as well as a strong subjective agreement on the hue values associated with these color labels. We further found that free recall for color was biased towards the mean hue value for each preferred color. We modeled this relationship between prior expectation and episodic memory with a rational model under the simple

assumption that people combine expectations for color with noisy memory representations. This model provided a strong qualitative fit to the data.

Introduction

Our knowledge and expectations learned from our environment shapes how we perceive, navigate, and interact with the world. They influence how we categorize objects and information (Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, & Vevea 2000; Jern & Kemp, 2013; Galleguillos & Belongie, 2010), how we visually perceive objects (Eckstein, Abbey, Pham, & Shimozaki, 2004; Epstein, 2008; Goldstone, 1995; Mitterer & de Ruiter, 2008; Todorovic, 2010), and how we make predictions (Griffiths & Tenenbaum, 2006). In memory, knowledge of the statistical regularities in the environment, such as the average height of people and the prototypical sizes of objects, exerts strong influences on how we recall such information (Bartlett, 1932; Hemmer & Steyvers, 2009a; Hemmer and Steyvers, 2009c; Hemmer, Tauber, and Steyvers, 2015; for a review see Hemmer & Persaud, 2014). Assuming that our expectations are environmentally derived, an important question for cognition is whether differences in environmental structure differentially influence expectations, and in turn episodic memory.

Color is one such feature that changes in representation across environments, and might engender differences in expectations. Individual and group differences in color knowledge and expectations have been attributed to communicative value (Meo, McMahan, & Stone, 2014), environmental occurrence (Stickles & Regier, 2014), and internal preferences (Palmer & Schloss, 2010). It has also been suggested that color category knowledge develops as a function of cultural experience (e.g. Roberson, Davies, & Davidoff, 2000). For example, there are significant differences in perceptual judgments for color between different cultural groups. This has been demonstrated in various

cultures including Russian, where there are two terms for blue (Paramei, 2005; Winawer, Witthoft, Frank, Wu, Wade, & Boroditsky, 2007), Papua New Guinea, who use 5 color categories (Roberson, Davies, & Davidoff, 2000), and a semi-nomadic South African tribe, who categorizes color based on light and dark (Roberson, Davidoff, Davies, & Shapiro, 2004). What remains to be examined is whether differences in the natural environment differentially influence long-term episodic memory across cultural and social groups.

The relationship between the structure of the environment and memory has been well described by Bayesian models of cognition (e.g., Shiffrin & Steyvers, 1997; Steyvers & Griffiths, 2008; Hemmer & Steyvers, 2009; Steyvers, Griffiths, & Dennis, 2006). This approach characterizes the computational problem people face when trying to recall real-world events under varying degrees of uncertainty. The models depict how an observer in a task integrates noisy and incomplete information stored in episodic memory with prior expectations for the environment when trying to recall an event. When the specific feature of an event is first experienced, this leads to noisy memory traces, centered on the original feature value, with some variation. It is also assumed that the observer has a prior expectation for the feature value that mirrors that of the distribution in the environment. The goal of the observer is to recall the feature value using noisy samples retrieved from memory and their prior expectation for the distribution of the feature value.

The assumption that memory is an integration of prior expectations with episodic traces stored in memory appears reasonable in the domain of color. For example, memory for color has been shown to be a blend of prior knowledge for object color typicality and episodic information (Belli, 1988). Belli found that reported color typicality of objects

(i.e. beverage pitchers were prototypically yellow) influences later color recognition. In his study, participants' recognition responses were a blend (i.e. yellow-green) of the actual study item (i.e. green pitchers) and prior knowledge (i.e. yellow pitchers). Similar findings result from a misinformation effect when post event information is blended with actual event information to produce recall (Loftus, 1977). Loftus found that recall for the color of a car was a blend (i.e. bluish-green) of the true color (i.e. green), and misleading information about the color of the car (i.e. blue).

To examine the influence of expectations learned from natural environments with different underlying representations of environmental features (e.g. color) on episodic memory, we conduct a cross-cultural investigation. Unlike previous research using simple memory measures to assess memory across cultures, such as percent correct (e.g. Roberson, et al., 2005), we characterize the optimality of the memory system and detail its relationship to the environment. We first quantify prior expectation for color in a standard U.S. undergraduate population. Prior expectations are assessed bi-directionally, both as a function of color naming preferences and the association of hue values to preferred color labels. Next, we employ a continuous recall task to assess the influence of prior expectation on recall for color. We implement a simple Bayesian modeling account to further characterize the relationship between expectations and episodic memory. Importantly, we contrast these findings with a cross-cultural study where we measure memory for color in an indigenous population whose natural environment is different than the standard U.S. population. We explore whether regularities in memory persists across natural environments or are dependent upon the different underlying representations for each environment.

Experiments

In experiments 1 and 2, we first sought to quantify peoples' bi-directional expectations for color, both as a function of color labeling preferences and the hue value associated with given color labels. The bi-directional assessment allowed us to examine linguistic categorization as well as category representativeness of color hue values. The resulting distributions over hue values were informative for the implementation of the Bayesian model (see section 3 Modeling). In experiment 3, we then assessed the influence of expectations on memory via a free recall color task. In all experiments, we collected data from as many individuals that volunteered to participant in the study.

Experiment 1: Color-Naming Task

Participants

Forty-seven Introductory Psychology undergraduate students at Rutgers University participated in this study in exchange for course credit. Data from one subject was discarded because no responses were recorded.

Materials and Procedure

The stimuli consisted of 48 colors sampled from the HSL (hue, saturation, luminance) color space. Colors varied in hue by 5 units (i.e. hue values of 0, 5, 10, etc) along the full hue range from 0-239, based on the ability to perceptually differentiate two sequential colors in the range. Saturation and luminance were held constant at 100% and 50%, respectively. A color patch measuring three-by-three inches was presented in the center of the computer screen. Participants were asked to provide a color label for that specific patch by typing their answer in a response box below the color patch. The patch remained on the screen until participants were satisfied with their responses and clicked 'continue'

to view the next patch. Each of the 48 color patches were presented twice in random order, for a total of 96 trials.

Results

Figure 1.1 shows label frequencies for the 48 hue values. The top panel shows the 7 most frequent labels (red, orange, yellow, green, blue, purple and pink). The 7 labels comprised 28% of all responses and coincide with the universal color terms of Berlin & Kay (1969). The bottom panel shows label frequencies for the top 21 labels, comprising 59% of total labels. The cutoff for including the 21 labels was based on a label being given a minimum of 40 times. The results show that participants expressed a large degree of agreement in the assignment of color labels to hue values. They also demonstrated a flexible color naming granularity for labels, with basic terms (e.g. red) and basic terms with modifiers (e.g. light green) being the most frequently used. This *suggests that participants have clear knowledge and expectations for color labels.*

Experiment 2: Color Generation Task

Participants

Forty-nine undergraduate students at Rutgers University participated for course credit or monetary compensation of $10. These participants were not involved in Experiment 1.

Materials and Procedure

The stimuli consisted of the 21 most frequent color labels given as responses in Experiment 1. The labels were presented one at a time, in 24 point Georgia font at the upper right side of the computer screen. The instructions were to generate the color hue that best corresponds to each of the labels using a color wheel. Color hue responses were generated by moving a cursor over a large black circle presented on the left side of the

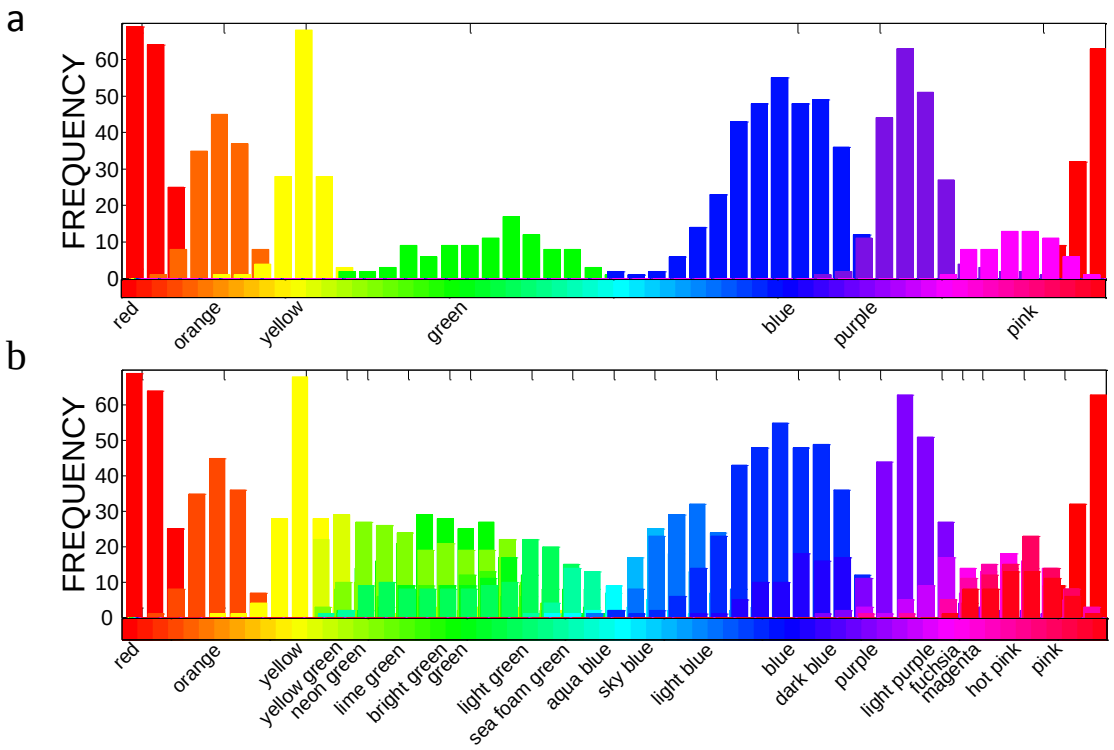

Figure 1.1 Frequency distributions over color labels in Experiment 1. (a) Frequency distributions over 7 most frequent labels. (b) Frequency distributions over 21 most frequent labels. Each bar represents a 5 unit range on the hue scale from 0-239.

computer screen. The black circle was a mask over a color wheel that varied in hue only. When the black circle was clicked, the corresponding color from that location of the underlying color wheel was shown in a three-by-three inch patch to the right of the wheel and below the color label. The underlying color wheel was rotated randomly by 45 degrees for each trial so that it was not possible to predict a color's location on the wheel from trial to trial. Participants were free to click as many times as they wished to generate the color they thought best corresponded to the given color label. Once participants were satisfied with the color they generated, they pressed the "space bar" to continue to the next trial. Participants generated colors for 21 labels twice each, for a total of 42 trials, presented in random order.

Results

The color wheel allowed participants to generate colors that differed by 1 unit of hue, resulting in 239 possible hue values. Responses where binned into 48 bins (varying by 5 units on the hue range from 0-239, such that all hue values that ranged between 2.5-7.5, were included in one bin, hue values between 7.5-12.5 fell in the next, and so on). Outliers more than 40 hue values from the highest or lowest value in a given colors hue range (see Table 1.1) may have reflected inattention to the task or accidental submission, and thus were removed, resulting in the removal of 11 responses (0.5% of the data). For subsequent model use, we fitted the frequency distributions with von Mises distributions (a.k.a. the circular analogue of the normal distribution). The means and standard deviations from the von Mises fits are shown in Table 1.1. Figure 1.2 shows frequency

Table 1.1. *Mean (SD) of Hue Values and Hue Ranges for Top 7 Color Labels*

	Mean (SD)	Hue Range
Red	1.1 (2.56)	(230–239, 0 – 5)
Orange	20.23(5.59)	(10-30)
Yellow	40.05 (3.04)	(35-50)
Green	79.79 (10.34)	(55-110)
Blue	153.53(12.13)	(115-170)
Purple	189.41 (6.27)	(175-190)
Pink	215.60 (9.57)	(195-225)

distributions over the hue values generated for the given color labels. The top panel shows the hue value frequency distributions for the 7 most frequent labels from Experiment 1 (red, orange, yellow, green, blue, purple and pink). Figure 1.2, bottom panel shows the frequency distributions for all 21 stimulus labels. The distributions reflect the notion that a given color label is best represented by a small range of hue

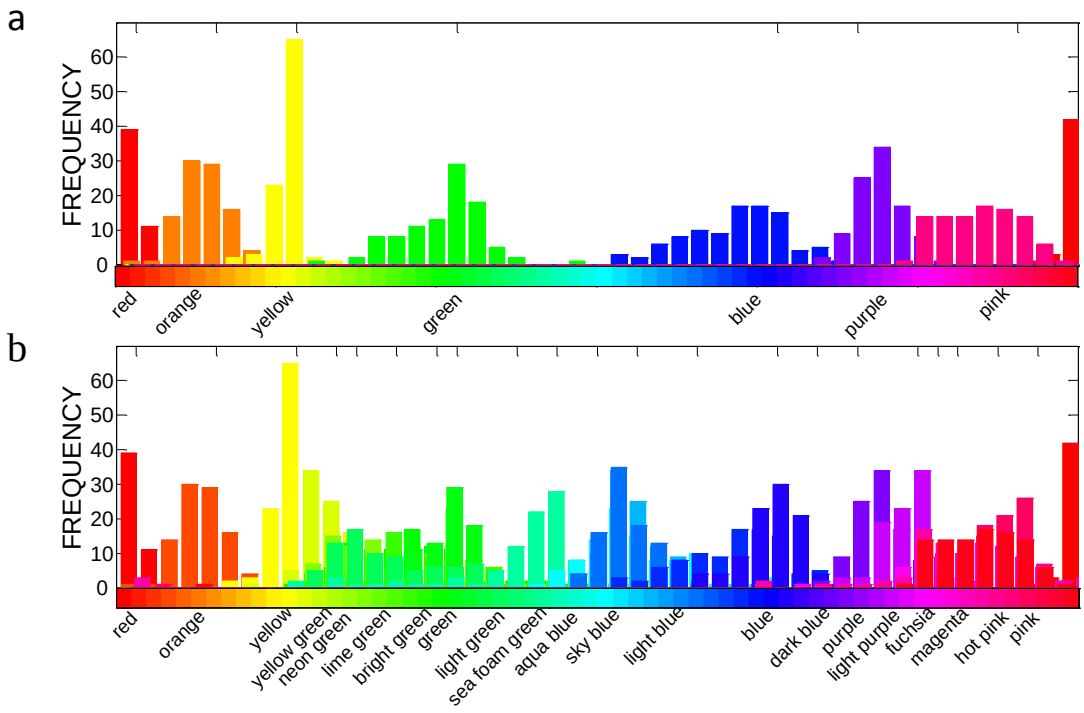

Figure 1.2. Frequency distributions over hue values from Experiment 2. Top panel: frequency at which a hue value was generated for the 7 preferred color labels. Bottom panel: frequency of hue values generated for the 21 most frequent labels. Each bar represents a 5 unit range on the hue continuum from 0-239.

values, with some overlap at the edges of the distributions. They also reflect strong agreement in the expectations for the association of color labels to hue values across participants.

Experiment 3: Color Memory Task

Participants

Eighteen Introductory Psychology undergraduate students at Rutgers University participated for course credit. These participants were not involved in Experiments 1 or 2.

Materials and Procedure

The stimuli consisted of 48 random shapes uniformly filled with the same 48 hue values in the HSL color space used in Experiment 1. Study and test trials were presented as a continuous sequence and were randomly interleaved (see Figure 1.3 sample study/test sequence). The color/shape pairings were randomized across participants and were presented one at a time, for 2 seconds each, at the center of the computer screen. On a test trial, a shape from a previous study trial, but filled with gray, appeared at the center of the screen and participants were asked to make three responses: 1) a recognition response: "do you remember studying this shape?" 2) a color label response: "What color was the shape at study?" (this question was posed regardless of their response to the recognition question). Responses were typed into a text box and participants pressed "enter" to continue. 3) a cued recall response for hue: "recreate the color of the shape at study". Responses were given using the same color wheel from Experiment 2[1] and were self-paced. Because of the continuous design where study and test trials were randomly interleaved, the lag between a study presentation and a test trial for that study stimulus varied from a lag of 1 to a lag of 48 (i.e., up to 47 intervening trials between study and test).

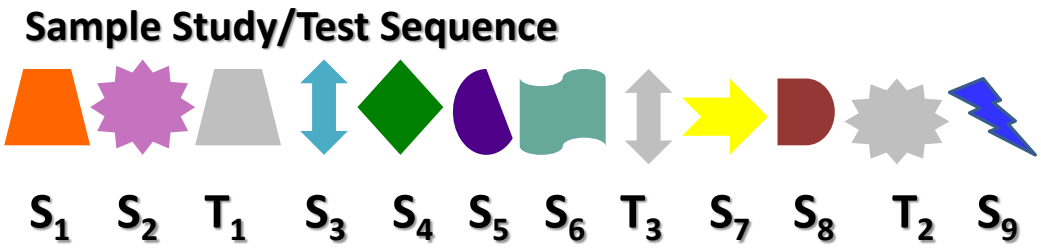

Figure 1.3. Sample study/ test sequence. S denotes a study trial and T denotes a test trial for with the trial number in subscript.

[1]To determine if the regression to the mean effect borne out in the memory data was merely a result of participants being primed by the label they recalled before recreating to color, we piloted another condition where participants recreated the color before providing a label, and the results mirrored the original memory condition.

Results

To measure performance, we calculated recall bias as the difference between the recalled and studied hue value. It appears that the task was very difficult, and error rates were very high. We therefore restricted the analyzed sample to include only cases in which subjects provided the correct label on the second question of the test trials (e.g. datum was excluded if the subject recalled blue, when the color studied was red (based on the most frequent label for that hue value in the color naming task), however, responses such as light blue, if the studied color was blue were acceptable). The hue range for a

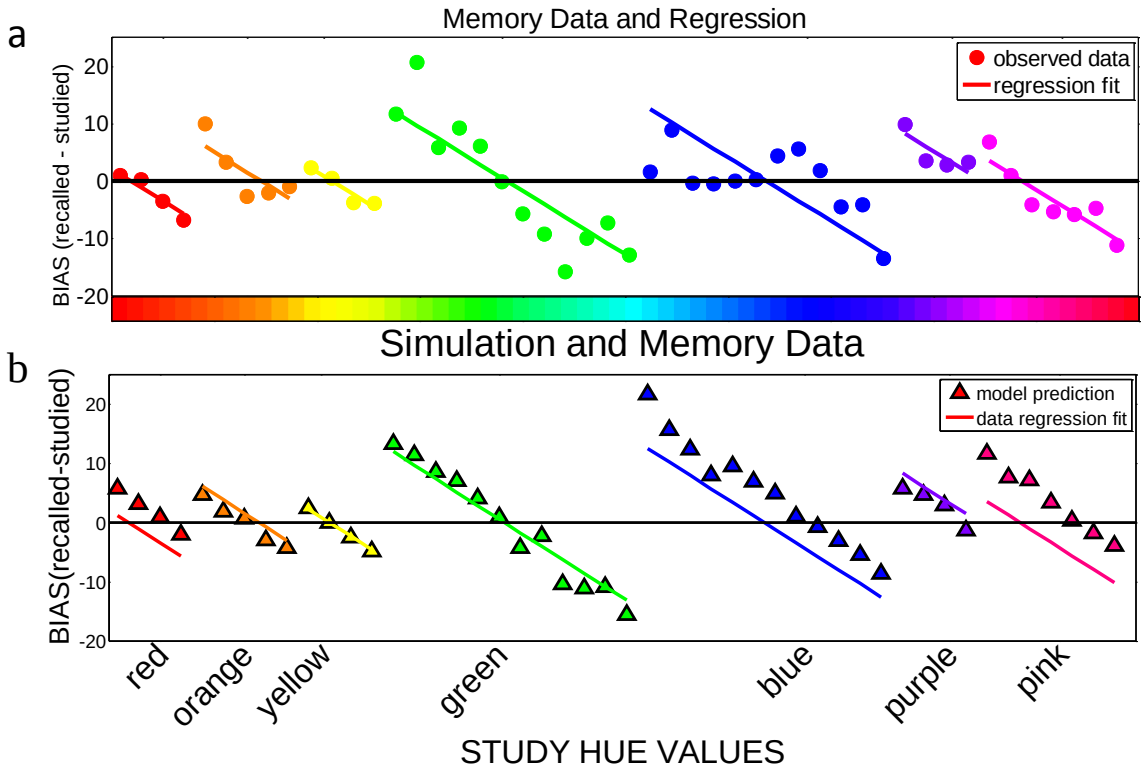

Figure 1.4. Top panel: Recall bias by color category. Positive bias indicates over estimation and negative bias indicates underestimation. The black line indicates no bias. The data points are color coded with the hue for that color range and the corresponding labels are given on the x-axis. The lines give the regression fits for each color label. Bottom panel: Model predictions with regression fits from the memory data.

color category was determined based on the lowest point between two response distributions in the color naming task. Furthermore, hue responses that deviated by more than 6 standard deviations from the mean of the determined hue range were excluded. This corresponded to correctly providing the label ‘blue’ to a blue hue value, but reconstructing it as red with the color wheel (4 data points). Five test trials were also excluded because no response was recorded. Thus, 55% of the data was used in this analysis.

The results revealed regression toward the mean effects as illustrated in Figure 1.4 top panel. For each of the 7 colors, subjects overestimated values below the mean hue value of each color category and underestimated the values above the mean hue of each color category. A linear regression model was fitted to each subject for each of the 7 preferred colors assuming a single slope and separate intercept for each regression line (see Figure 1.4 top panel). A one-way analysis of variance revealed a significant main effect of intercepts ($F[694]=664$, $p<.001$) across color categories. The negative slope of the lines indicates a regression to the mean, and the different intercepts for each of the color categories signify regression towards different mean values. Table 1.2 shows the slope and intercepts for the 7 categories.

Modeling

In this section we implement a simple Bayesian model to characterize the regression to the mean effect borne out in the memory experiment. In the model, the goal is to efficiently retrieve relevant information from memory, which needs to be combined with prior knowledge and expectations about the environment. Bayes’ rule gives a principled

account of how to combine noisy memory representations with prior expectations to calculate the posterior probability,

$$p\,(\Theta|y) \propto p\,(y\,|\Theta)\;\; p\,(\Theta) \qquad \text{Eq (1)}$$

where the posterior $p(\theta\,|y)$ gives the likely feature value θ given the noisy memory content y. We assume that the studied features (i.e., hue values) are Gaussian distributed, $\theta \sim \mathrm{N}\,(\mu,\, \sigma^2)$, with the prior mean μ and variance σ^2 of the features drawn from the environment. When the specific feature θ is studied, we assume this leads to memory traces y, with some memory noise ψ, $y \sim \mathrm{N}(\theta,\, \psi)$. Standard Bayesian techniques (Gelman et al., 2003) were used to compute the mean of the posterior distribution:

$$\hat{\theta} = w\mu + (1 - w)\bar{y} \qquad \text{Eq (2)}$$

where $w = (1/\sigma_0^2)/\,[(1/\sigma_0^2) + (\mathrm{n}/\sigma_m^2)]$ and n is the number of samples taken from episodic memory.

We specified a prior with mean μ for each color category equal to the mean of the von

Table 1.2. *Mean Slopes and Intercepts by Color Label*

	Slope		Intercept	
	Mean	SD	Mean	SD
Red	-0.46	0.13	-3.4137	3.26
Orange	-0.46	0.13	10.6451	2.63
Yellow	-0.46	0.13	18.2125	3.61
Green	-0.46	0.13	37.0389	8.79
Blue	-0.46	0.13	64.9861	4.43
Purple	-0.46	0.13	88.1715	4.90
Pink	-0.46	0.13	92.4914	6.79

Note. N=18

Mises (circular analogue of the Gaussian) distributions calculated from the data in Experiment 2. In other words, we assume these distributions to be representative of peoples' prior expectation over hue values for a given color category. In the same way, we set σ^2 for each color category equal to the variances of those distributions and a memory noise (ψ) that varies for each category on the standard deviations from those distributions from Experiment 2 (see Table 1.1). While Bayesian cognitive models are generally hand-fitted to the data, here all parameter settings are directly informed by the experimental data. We used the model to simulate the same trials in the experiment. Figure 1.4 bottom panel shows the simulated responses from the model. Overall, the model produces results that are qualitatively similar to the observed data and captures the overall trend. This provides strong support for reconstruction from memory being highly systematic and influenced by prior expectations learned from the environment. Next, we build on this principle of systematicity, that we assume is a fundamental mechanism of memory, to investigate how different environments and potentially different expectations for color might influence regression patterns in memory.

Discussion

In this work we sought to investigate the influence of expectations for color on episodic memory. We measured prior expectation via two tasks: a color naming task which elicited color naming preferences, and a unique task in which participants used a color wheel to generate colors most closely associated with the given color label. The results showed naming preferences that are consistent with the existing literature (Berlin & Kay, 1969), namely red, orange, yellow, green, blue, purple and pink. Subjects also showed a high level of agreement in both Experiments 1 and 2. We then measured the influence of

expectation on free recall for color. Results revealed a regression to the mean effect in free recall, such that studied hue values below the mean of that color category were overestimated at recall and studied hue values above that color category were underestimated. This suggests that recall is influenced by expectations for color.

This behavior was modeled with a simple rational model of memory, which assumes that prior knowledge for different color categories exert an influence on episodic recall. In this way, recall is a combination of prior expectations and noisy memory content. The model provides qualitative predictions that are a good fit to the observed data. The model captures the regression to the mean effect for each of the 7 preferred labels. Importantly, the only assumption made in the model was that prior expectations for color were well described by the performance in the color generation task.

Here, we do not provide an analysis of sub-labels (all 21 labels). However, results for hue values within the blue range are interesting in that the pattern of over and underestimation appears to be dispersed. This may be the result of participants separating the hue values in the blue range to account for not just the universal label ‘blue’, but also high frequency sub-labels (i.e. light blue and sky blue). This suggests that colors might be hierarchically organized, such that blue is the general color label, and sub-labels are based on subjective naming preferences. We believe that this investigation has provided important support for existing understanding of the structures of color categories, as well as a new understanding of relationship between prior expectations and free recall for color.

Chapter 3: Inferring Prior Knowledge from Episodic Memory in Special Populations

This chapter presents data from a study previously published in the *Proceedings of the Annual Meeting of the Cognitive Science Society* and in *i-Perception.* K. Persaud and the advisor, P. Hemmer, developed the study concept and study design together. K. Persaud developed the stimulus. C. Kidd and S. Piantadosi, performed the testing and data collection which was conducted in a different country (Bolivia). K. Persaud performed the data analysis. K. Persaud and P. Hemmer, together, performed the interpretation. K. Persaud drafted the manuscript. After the manuscript was drafted, all authors helped revise the manuscript. K. Persaud implemented all critical revisions in response to reviewer comments.

In the study, we sought to examine memory in a population that might have dissimilar expectations from our standard US population based on their natural environment and culture. These expectations in turn might differentially influence memory. We engage this question in the domain of color for a number of reasons. Color holds social and cultural relevance and people's relationship to color can be both internally (e.g. emotional connections to color) and externally (e.g. through the visual experience in their environment) derived. In addition, color is a ubiquitous domain for research across developmental, social, and cultural groups, as well as across domains of cognition.

Importantly, for investigative purposes people have similar, but also different knowledge states of color. There is an extensive literature characterizing knowledge of color across cultures (e.g., Davies & Corbett, 1997; Regier, Kay, & Cook, 2005; Roberson, Davidoff, Davies, & Shapiro, 2004; Stickles & Regier, 2014; Xu, Griffiths, & Dowman, 2010), and several clear patterns of color universality have emerged. For

example, it has been shown that universal tendencies persists in color naming across societies (Berlin & Kay, 1969; Regier, Kay, & Cook, 2005) and that those tendencies are linked to 11 basic color terms (i.e., red, orange, yellow, green, blue, purple, pink, black, white, gray and brown). A possible source of universal tendencies in color naming is similarities in favored color percepts (i.e. best examples) across various languages (Regier, Kay, & Cook, 2005). These color universals are shown to have a subjective perceptual basis, in that they can be used to partition the color space into distinct regions that facilitate color categorization (Webster & Kay, 2012).

While these 11 universal categories are found across most industrialized societies, there are also substantial individual, environmental, and cultural differences in color knowledge (e.g., Palmer & Schloss, 2010; Stickles & Regier, 2014). Internal (e.g., emotional) relationships and preferences to certain colors serve as a candidate source of variation in individual color knowledge as postulated by the Ecological Valence Theory of Human Color Preferences (Palmer & Schloss, 2010). This theory posits that people's emotional response to a color is their cumulative affective response to the objects to which the color is associated. Individuals prefer colors that they have had positive experiences with (e.g. yellow – color of flowers) and do not prefer colors with which they have had bad experiences (e.g. red – color of fire), signifying each person's close and personal relationship to color.

At the group level, a source of variation in subjective color knowledge is the relationship between color and the variability in natural environments. For example, color terms in languages with climates of abundant vegetation (e.g. rainforest) are significantly different from color terms in languages with dry climates (e.g. Savanna), but not in places

with relatively similar climates (e.g. rainforest and monsoon) (Stickles & Regier, 2014). The difference in the greenery of the climates presumably accounts for difference in color naming. Thus, it appears that local environmental factors influence color knowledge and promotes variability in color terms across languages.

We tested recognition memory for color in the Tsimane′ group of Bolivia. The Tsimane′ are an indigenous people who inhabit rainforests east of the Andes in lowland Bolivia. They have minimum contact with the outside world, a uniquely different color diet relative to our U.S. population, and varying levels of education (see table 2.1). These factors might contribute to idiosyncratic expectations for color. Furthermore, the difference in expectations may be foreshadowed by dissimilarities in color language. In the Tsimane′ language, color terms are highly variable and morphologically complex—e.g., yellow is called "color-of-the-cuchi-cuchi-tree". Color language is also inconsistent in that some people know this term for yellow, as well as other color terms, and some do not.

Color expectations of the Tsimane′ people may lead to three possible regression patterns. 1) The pattern might be the same as the U.S. population, such that memory regresses to the same seven color categories, suggesting that the two populations used the same categories regardless of environmental variation. 2) The patterns of the two populations might differ, in that the Tsimane′ could potential combine some color categories. This is supported by smaller numbers of color categories across some languages (e.g., Roberson, Davies, & Davidoff, 2000; Roberson, Davidoff, Davies, & Shapiro, 2005). 3) The Tsimane′ might split some categories–e.g., as observed in Russian where blue has two terms (e.g., Paramei, 2005). Such a split could be based on the high

Table 2.1. *Participant Demographics*

Age (years)	18	20-28	30-34	40-48	60+	
Frequency	4	8	6	3	2	
Education (years)	0	1	2	3-5	6-9	10
Frequency	4	1	3	9	5	1
Spanish (translate out of 11)	0	6-9	10-11			
Frequency	1	19	3			
Counting (highest #)	2	5-9	15-31	46-64	93	102
Frequency	1	2	5	3	1	11
Arithmetic (out of 12)	0	1	2-3	4-5	6	10-11
Frequency	2	3	10	2	2	3

Note. N=23

variability in color terms in the Tsimane′ language, and their natural environment. Regression towards the standard universal color categories in both populations would suggest that these factors (language variability and environment differences) may have little influence on memory. Alternatively, differences in regression patterns would provide support for cultural and environmental factors influencing memory.

Due to the demands of field research, the task varied in a number of ways compared to the controlled laboratory experiment. First, the Tsimane′ displayed a great deal of discomfort with the use of technology and any apparatuses that they themselves had to use. Thus, we converted from a computerized free recall task to a paper based recognition task where participants only needed to point to responses. Second, instructions and responses required two layers of translation (i.e. from English to Spanish, and then from Spanish to the Tsimane′ language), and thus we were unable to assess prior knowledge and expectations as was previously done with the U.S. population. We instead relied on the systematicity of memory (i.e. regression to the mean effect) and the assumptions of

the Bayesian cognitive model to infer the underlying color categories of the Tsimane′ and the influence on memory.

Experiment: Episodic Memory for color in an indigenous population

Participants

Twenty-three individuals participated in this study and were compensated with small gift bags of local goods. Participant ages ranged from 18-65. Self-reports of education levels ranged from no formal education to 10 years of education, and arithmetic skills ranged from 0-11 out of 11 questions correct on an ad hoc field measure (using all addition questions), and highest count ranging from 2-102 (meaning knowing all numbers). Table 2.1 gives a detailed breakdown of the demographics and skill variables.

Materials and Procedure

Stimuli consisted of 24 random shapes uniformly filled with 24 unique colors sampled from the hue color space, with saturation and luminance held constant at 100% and 50%, respectively. The 24 colors were selected from the 7 color categories and varied in hue by a minimum of 5 units (on a total range of 239). Furthermore, colors were randomly selected from each color category, proportional to the size of the color category (i.e. 2 red, 3 orange, 2 yellow, 6 green, 6 blue, 2 purple, and 3 pink). Study shapes were printed individually, and test shapes along with 5 distractors, were printed together on 5.5-by-8 inch cards (See Figure 2.1a for a sample study test pair). The colors of the distractors were chosen such that the hue values of two distractors were greater than the hue value of the target color, two distractors were less than the hue of the target, and the last distractor hue value was either greater or less than the target, but at a further absolute distance from the target than the other distractors (see Figure 2.1a for illustration).

Participants were gathered in a communal classroom, and there were a number of onlookers during the administration of the test. Figure 2.1b shows both the experimental setting and a study-test trial sequence. A translator explained the task, and all participants appeared to immediately understand the procedure. Presentation time of the 1-item study card was as close to 1 second as possible. The study trial was followed by 6-alternative forced choice immediate recognition. Participants had as much time as they needed, but most responded immediately, and responses were recorded in a booklet. On some trials

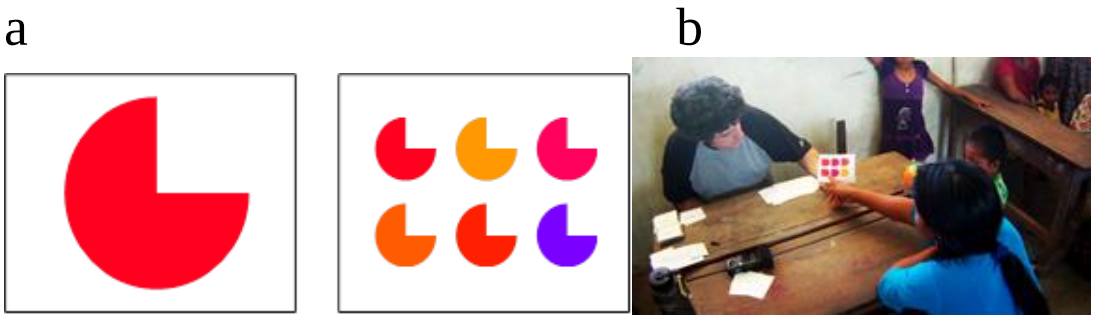

Figure 2.1. (a) Sample study-test stimulus. (b) Tsimane' woman participating in the study. The experiment was conducted in a class with onlookers from the community.

(approx. 5%) it was not clear where the participant had pointed, and participants were asked to repeat their choice. They were asked to touch, rather than point, to try to alleviate this problem. Trial order was randomized between participants. Due to the field demands, it was not possible to randomize the target/distractor locations on the test trials. This means that all participants saw identical test cards.

Results

Prior to analysis, recognition responses that were more than 6 standard deviations away from the studied hue value were removed. These data points constituted 2.5% of all the data (14 out of 545 data points). After calculating the bias measure described below,

individual subject data revealed that there was one participant whose data were very noisy and appeared essentially random (this was not unexpected given the very noisy conditions of field data collection). This may have reflected either impairment in color vision[2] or inattention to the task and this participant's data was removed from all further analysis.

Recognition Bias and Regression Memory performance was measured in terms of recognition bias, i.e., the difference between the hue value participants remembered and the hue value studied. First, bias was calculated for each individual participant and then averaged across participants for each studied hue value. Figure 2.2 shows recognition bias as a function of studied hue values. The data show clear regression to the red, green, blue and pink color categories. The orange, yellow, and purple categories, however, were more ambiguous. Based on a visual inspection, we partitioned the averaged bias into 5 categories—combining orange and yellow, and combining purple and pink—and fit a linear regression model to each of the 5 resulting color categories (see Figure 2.3). The slope of the regression in each category (except for the orange/yellow range) was negative, indicative of a standard regression to the mean effect. Hue values below the mean of the category were overestimated and hue values above the mean were underestimated. This is consistent with the findings from experiment 3. A one-way analysis of variance revealed a significant main effect of intercept ($F[109]=25$, $p<.001$) across color categories, indicating that each category has a different intercept. However, performance in the orange/yellow range appeared to be different from the other

[2] We were not able to conduct a color blindness test. The assessment requires naming knowledge of some shapes which is confounded with education. Many Tsimane′ participants could not complete this task.

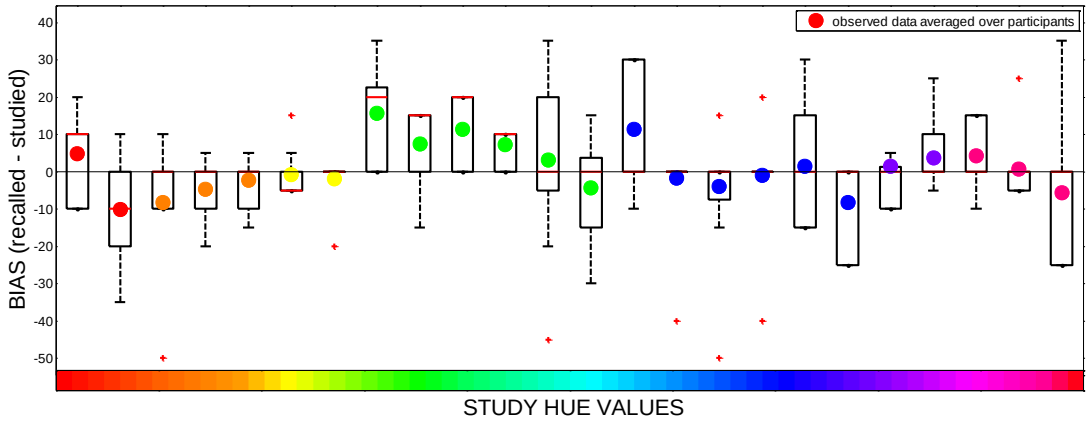

Figure 2.2. Recognition bias by hue value. Average mean bias (data points) and response ranges (box plots) for each studied hue value. Colors of the data makers indicate the standard universal color categories. Positive bias indicates over estimation and negative bias indicates underestimation. The black line indicates no bias.

categories. In this category, the slope ran in the opposite direction (positive slope), showing a regression towards orange-red rather than towards yellow.

Cluster Analysis Figure 2.3 appears to show interestingly different color categories, compared to the seven classic basic color terms (red, orange, yellow, green, blue, purple and pink). To learn the underlying categories that participants may have used, we conducted a k-means cluster analysis (Figure 2.4). We ran 10 iterations of the cluster analysis on four different clusters sizes (i.e., 4, 5, 6, and 7) and found the greatest cluster agreement over the 10 chains for a cluster size of 5. This cluster size was further confirmed by the Calinski Harabasz criterion. Consistent with the regression analysis, the cluster analysis also combined colors in the purple/pink ranges and orange/yellow ranges. However, the cluster analysis further combined the orange/yellow category with red, but split the universal blue range into two blue categories. These findings suggest that the pattern of regression behavior to underlying category centers is inherent to memory, but

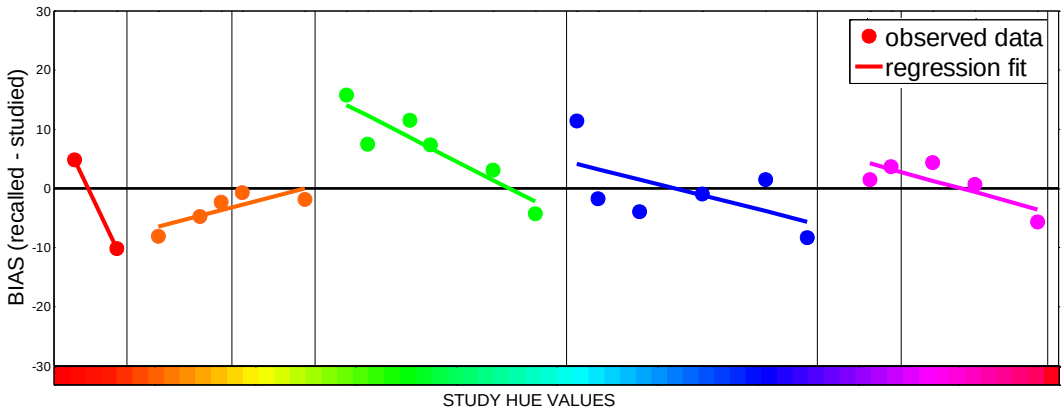

Figure 2.3. Regression fits to 5 color categories. Categories are partitioned by hue ranges with orange and yellow combined, and pink and purple combined. The thick center black line indicates no bias. The data points are color coded with a hue for that color category. The lines give the regression fits for each of the 5 categories.

the specific categories— either assessed experimental (U.S subject population), or learned from the cluster analysis (Tsimane′)—are environment dependent, and are reflected in the differential regression behavior between the two subject populations.

Discussion

We examined expectations for color and the influence of those expectations on episodic memory in two populations: a standard U.S. population and the Tsimane′ people of Bolivia. We found that environment appears to differentially influence category expectations, and episodic memory. In the U.S. subject population, expectations reflected naming preferences that were consistent with the existing literature (Berlin & Kay, 1969), and a high level of subject agreement on the association of labels to hue values. Furthermore, in this population recall regressed toward 7 color categories, suggesting an influence of expectations for color categories on episodic memory.

In previous work, we modeled this relationship between expectations and memory with a Bayesian cognitive model characterizing the computational problem of combining prior expectations and noisy episodic content. Importantly, the only assumption made in the model was that prior expectations for color were well described by the performance in the color generation task. We believe this reflects the optimality of the memory system and its relationship to the environment. This gives rise to the question of whether different environments, cultural profiles (such as language), or experiences engender variation in color expectations and lead to differences in regression behavior.

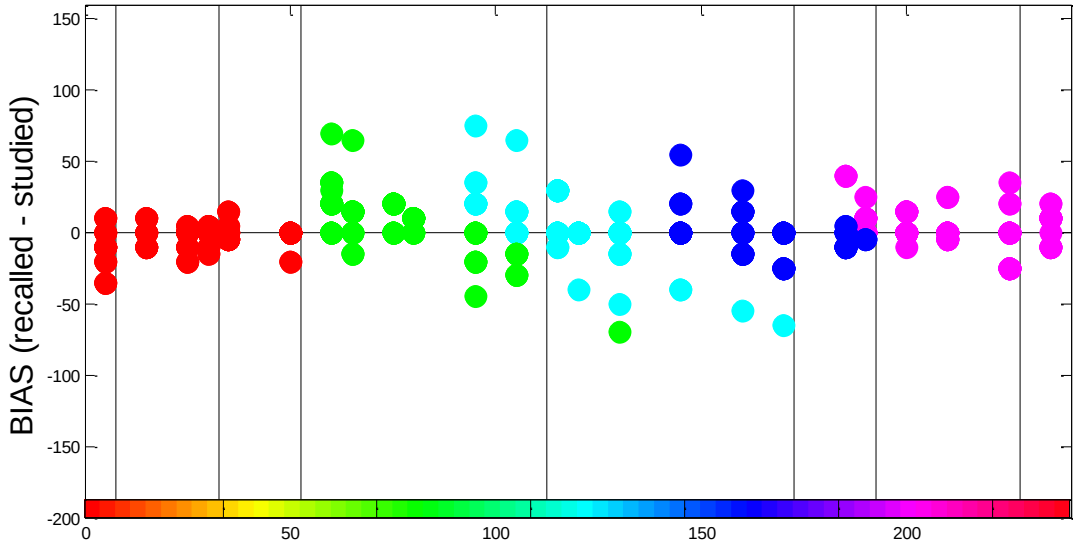

Figure 2.4. K-means Cluster Analysis. Bias data partitioned into 5 learned clusters from an unsupervised k-means cluster analysis, and color coded with a hue from that category. Vertical lines and color labels on x-axis show the standard universal categories.

To examine whether different environments engender variation in color expectations and lead to differences in regression behavior, we assessed memory in an indigenous population, the Tsimane′ of Bolivia. Due to field work constraints, we were unable to assess prior expectations for color or utilize the free recall memory design with this group. Instead, we worked backwards using the Bayesian assumption of the influence of

expectations on memory and the results of the U.S. memory experiment to learn the underlying categories for this group, and in turn, how these category expectations impact memory performance. In this work, two clear patterns emerged. We found a consistent regression to the mean effect across color categories, with the exception of the yellow category. This finding may suggest that the regression to the mean effect in memory is a universal cognitive process and is systematic across cultural and environmental groups. Interestingly, however, a k-means cluster analysis showed that the categories in the Tsimane′ population were different than observed in a standard U.S. population. While the U.S. group regressed toward seven categories, the Tsimane′ segregated blue into two categories, and combined other categories, resulting in five inferred categories: red/orange/yellow, green, light blue, dark blue, and purple/pink.

The population specific bias observed in the Tsimane′, relative to a U.S. population, might be related to the underdevelopment of knowledge for some categories. This could be due to one or more factors, such as low environmental incidence, low frequency in language, limited formal education of color, or little communicative need of certain color terms. From a memory perspective, the underdevelopment of color categories raises several interesting questions. A color like yellow, which is somewhat rare in the Tsimane' environment, might lead to an outlier (or Von Restorff) effect, where it is better remembered. Conversely, a pervasive color (with a high prior probability in the environment) is also likely to lead to better memory, and might account for the shallow regression line in the blue category (Figure 2.3).

We believe that this study provides important evidence for an experience based mechanism (development and maintenance of prior knowledge) that gives rise to

differences in color knowledge. This is consistent with the findings of Stickles and Regier (2014) that environment impacts language (i.e., color words). Furthermore, the study provides strong support for the influence of category knowledge on memory, and the systematicity of memory across groups with varying prior knowledge content.

Chapter 4: Fidelity and Memory

The Dynamics of Fidelity over the Time Course of Long-term Memory

Kimele Persaud and Pernille Hemmer (2016). *Cognitive Psychology*

K. Persaud and the advisor, P. Hemmer, developed the study concept and study design together. Stimulus creation, testing and data collection were performed by K. Persaud. K. Persaud performed the data analysis. Interpretation of the analysis and model development/implementation was performed by K. Persaud and P. Hemmer, together. K. Persaud drafted the manuscript. After the manuscript was drafted, K. Persaud and the advisor, P. Hemmer, revised the manuscript. K. Persaud implemented all critical revisions in response to reviewer comments.

Abstract

Bayesian models of cognition assume that prior knowledge about the world influences judgments. Recent approaches have suggested that the loss of fidelity from working to long-term (LT) memory is simply due to an increased rate of guessing (e.g. Brady, Konkle, Gill, Oliva, & Alvarez, 2013). That is, recall is the result of either remembering (with some noise) or guessing. This stands in contrast to Bayesian models of cognition which assume that prior knowledge about the world influences judgments, and that recall is a combination of expectations learned from the environment and noisy memory representations. Here, we evaluate the time course of fidelity in LT episodic memory, and the relative contribution of prior category knowledge and guessing, using a continuous recall paradigm. At an aggregate level, performance reflects a high rate of guessing. However, when aggregate data is partitioned by lag (i.e., the number of presentations from study to test), or is un-aggregated, performance appears to be more complex than

just remembering with some noise and guessing. We implemented three models: the standard remember-guess model, a three component remember-guess model, and a Bayesian mixture model and evaluated these models against the data. The results emphasize the importance of taking into account the influence of prior category knowledge on memory.

Introduction

An important question for memory is whether category knowledge biases performance, and whether an influence of category knowledge changes as a function of the fidelity of memory. Recent work in visual working memory has suggested that when recalling stimulus features, observers either remember the episodic information with some noise or guess (Brady, Konkle, Gill, Oliva, & Alvarez, 2013; Zhang and Luck, 2008). Zhang and Luck found that fidelity is fixed once capacity of visual working memory is reached, but that the guessing rate changes. The resulting error distributions are well fit by a mixture of a Gaussian-like (remembering with some noise) and uniform distribution (guessing). They argued that observers remember continuous feature values and are not biased by categorization of those values. Importantly, a finding of category bias would suggest an intermediating step between remembering and random guessing. Such a bias was found by Bae and colleagues, establishing that category biases originate in perception and are reflected in visual working memory (Bae, Olkonnen, Allred, & Flombaum, 2015).

Several extensions to the original remember-guess model have been implemented to account for additional factors that influence visual short-term and working memory

performance (e.g., Bays, Catalao, & Husain, 2009; Bays, Wu, & Husain, 2011; van den Berg, Shin, Chou, George, & Ma, 2012). For example, the variable-precision model (VP; van den Berg, et al, 2012) postulates variability in the precision with which items are encoded in working memory. The resulting error distribution is a mixture of many von Mises distributions (as opposed to the one memory component in the remember-guess model), to account for residual noise in memory that the standard model cannot fit. Other proposed models incorporate task-based components, such as "misassociation" or "misbinding" parameters to extend the standard remember-guess model (Bays, Catalao, & Husain, 2009; Bays, Wu, & Husain, 2011).

Although these models provide substantial revisions to the original, it is important to note that they are grounded in visual short-term and working memory. Relatively few studies have sought to apply the remember-guess framework to understanding long-term episodic memory. One such application by Brady and colleagues (2013) showed that there is a loss of fidelity from working into long-term (LT) memory. They argued that this decrease in fidelity is due to an increased rate of guessing, without addressing other factors that impact long-term memory.

The remember-guess model stands in direct contrast to a number of Bayesian cognitive models which assume that LT memory is an integration of expectations learned from the environment with noisy memory representations (e.g., Hemmer & Steyvers, 2009; Hemmer, Tauber & Steyvers, 2015; Hemmer, Persaud, Kidd, & Piantadosi, 2015). These models are pervasive in cognition in general, and in specific domains including categorization (e.g., Huttenlocher, Hedges & Vevea, 2000), generalization (e.g. Griffiths & Tenenbaum, 2006), semantic memory (Hemmer & Steyvers, 2009b; Steyvers,

Griffiths, & Dennis, 2006), and episodic memory (Shiffrin & Steyvers, 1997; Steyvers & Griffiths, 2008).

Bayesian models of cognition propose a tradeoff between the fidelity of memory content and the influence of prior expectations. When the fidelity of the episodic trace is high, for example, as in visual short-term memory, there is minimal noise and potentially little influence of prior expectations. As fidelity decreases in working and LT memory, whether as a function of time or errors in retrieval, the influence of prior expectations would increase.

At an aggregate level, however, the error distributions resemble a combination of precise and imprecise memory, which might appear only to be remembered content and guessing, effectively masking underlying stages between the two. Prior expectation is a potential factor that might compensate for decreasing memory fidelity at the stage between precise memory and random guessing. In point of fact, Donkin and colleagues (2014) showed model-based evidence from visual short-term memory positing three discrete states of memory: One, a state based on perceptual memory and high precision, two, due to memory decay from perception, a state with intermediate precision based on verbal labeling, and three, guessing. Here, we seek to compare the performance of models that have been employed to characterize long-term memory, namely the remember-guess model (Brady et al, 2013) and Bayesian models of long-term memory (e.g., Hemmer & Steyvers, 2009; Persaud & Hemmer, 2014).

In the present work, we explore what happens to the precision of memory over time. Partitioning performance by the number of intervening trials between study and test (i.e., lag) allows for the systematic assessment of the time course of fidelity in LT episodic

memory. To the best of our knowledge, this paper gives the first analysis of free recall by lag in an effort to understand the relative contributions of prior knowledge and guessing. We also investigate if category bias, indicative of the employment of prior knowledge, is a mechanism by which LT memory can be filled in, before individuals resort to random guessing. If this is the case, then performance at intermediate lags, consistent with the Bayesian assumption, should reflect the influence of category knowledge on noisy episodic representations. Such an influence is generally observed as a regression to the mean effect. We implement three models: the standard remember-guess (RG) model, a three component remember-guess (3CRG) model, which assumes two levels of precision in memory and a Bayesian mixture (BM) model. We also conduct model comparisons as a function of lag.

Memory for color: Overview of Experiment

Our objective was to determine the contribution of prior expectations to LT episodic memory and assess the resulting time course of errors. We developed a novel experimental approach for assessing free recall for color, where participants generated recalled hue values using a continuous color wheel, and with interleaved trials of random lag lengths between study and test.

Participants

Sixty-one Introductory Psychology undergraduate students at Rutgers University participated for course credit or $10 compensation. In condition 1 (Label first Condition) N=18. In condition 2 (Label after Condition) N=5. In condition 3 (No Label Condition) N=38. All participants reported having normal color vision. No individual participated in more than one condition.

Materials

The stimuli consisted of 48 arbitrary shapes uniformly filled with 48 colors sampled from the winHSL240 (hue, saturation, and luminance) color space. See Figure 3.1 for sample stimuli. The shapes were selected such that there was little prior association of any color to the shapes, that is, the study set did not result in canonical pairings such as yellow stars or red hearts. The purpose of the shapes was to cue subjects on test trials to recall the fill-color of the shape. Colors were sampled in 48 equally sized steps along the full hue range, based on the ability to perceptually differentiate two sequential colors in the range. Saturation and luminance were held constant at 100% and 50%, respectively. The shapes and colors were paired randomly, and pairings were randomized across participants. Each shape and color was studied only once.

Procedure

Participants were shown a continuous study-test sequence of color filled shapes.

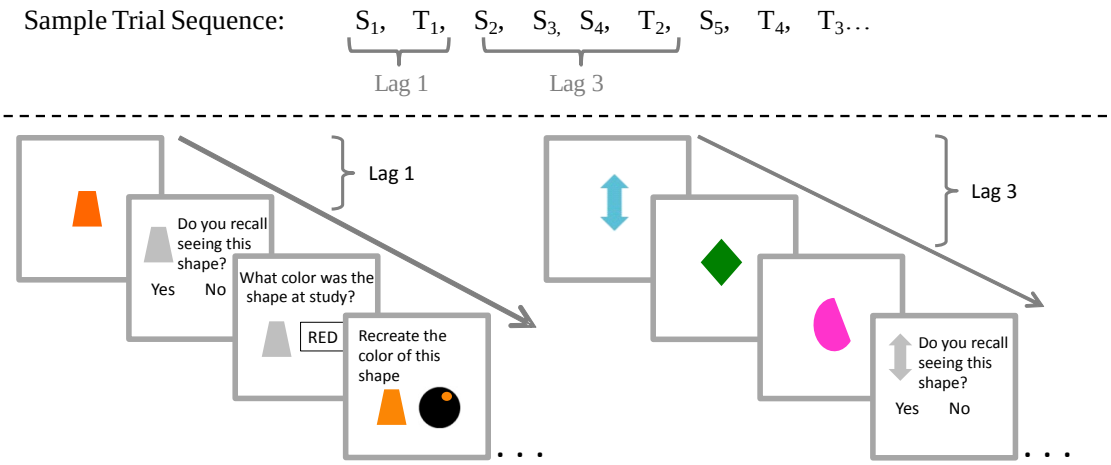

Figure 3.1. Sample study/test sequence by lag. Lag 1, participants study a shape, followed by a series of memory questions related to the color of the shape. Lag 3, participants study a sequence of three colored shapes, before being asked a series of memory questions related to the color of the cued shape – here, the first of the three shapes studied.

Shapes were presented one at a time at the center of the computer screen for 2 seconds. Participants were told to study the color of each shape, as they would be asked to recall the color of the shapes. Test trials were randomly interleaved between study trials, resulting in lags of varying length. This sequence of lag was obtained by first randomly permuting the order of study trials, and then interleaving test trials, with the condition that for a test trial to occur, the corresponding study item must have occurred first. Figure 3.1 provides an example of the experimental procedure for a lag of 1 and a lag of 3 trials, as well as an illustration of the interleaved study test sequence.

On a test trial, a shape from a previous study trial, but filled with gray, was presented as a cue and participants were prompted to make several responses. In all three test conditions, participants first completed a recognition task for the shape. In the two label conditions, participants were asked to provide a verbal label for the color of the shape either before or after recreating the shape color (this question was posed regardless of their response to the recognition question). Participants typed responses into a text box and pressed "enter" to continue. In condition 3, participants did not provide a verbal label. In all three conditions, participants were then asked to recreate the studied color of the shape using a continuous color wheel. The color wheel was covered by a black mask, and was randomly rotated by 90 degrees on every test trial. Participants clicked on the wheel to fill the shape with the underlying color. Test trials were self-paced.

Results

For analysis, and to accommodate the use of von Mises distributions in the models, hue values were converted from the winHSL240 color space to degrees.

The primary purpose of the three labeling conditions was to check that the explicit label generation did not alter the influence of category knowledge. We find no real differences between the label versus no-label conditions, and for the purposes of analysis, data is pooled across all three conditions (see Appendix Table A1 for parameter estimation for the label versus no-label conditions).

Lag Analysis To measure the time course of fidelity in LT memory, the data was partitioned by lag and each resulting error distribution was analyzed. Since lag intervals encompassed participant responses which were self-paced, lag intervals varied both across trials (with the same lag) and across participants. For an approximation of the correspondence of lag intervals to units of time, we calculated the average study plus response time for each condition and collapsed across conditions. The results were as follows: Label First: M= 16.2s, SD= 9.0s, MO= 10.0s; Label Last: M= 18.5s, SD= 8.2s, MO= 10.0s; No label: M= 11.2s, SD= 7.1s, MO= 5.0s; All conditions: M= 13.3s, SD= 8.3s, MO= 9.0s. A Pearson's correlation revealed a strong positive correlation between lag and response times ($r = 0.7$, $p<.000$).

Initially, all lag groups were examined separately, but then grouped based on a meaningful progression in the parameter contributions. This was done both for visual clarity, and in order to increase the "speed" of model fitting. See Appendix Table A2 for fits to all lags. Figure 3.2, from left to right shows the error distributions for lag 1, 2-3, 4-9, 10+, and the aggregate of all lags. The error distributions reveal that the fidelity of memory is quite high at lag 1. This is evidenced by the tight grouping of responses around 0 error and virtually no responses past 50 degrees of error. For the remaining lags, memory fidelity is not as high as in lag 1, but does appear to be stable over time.

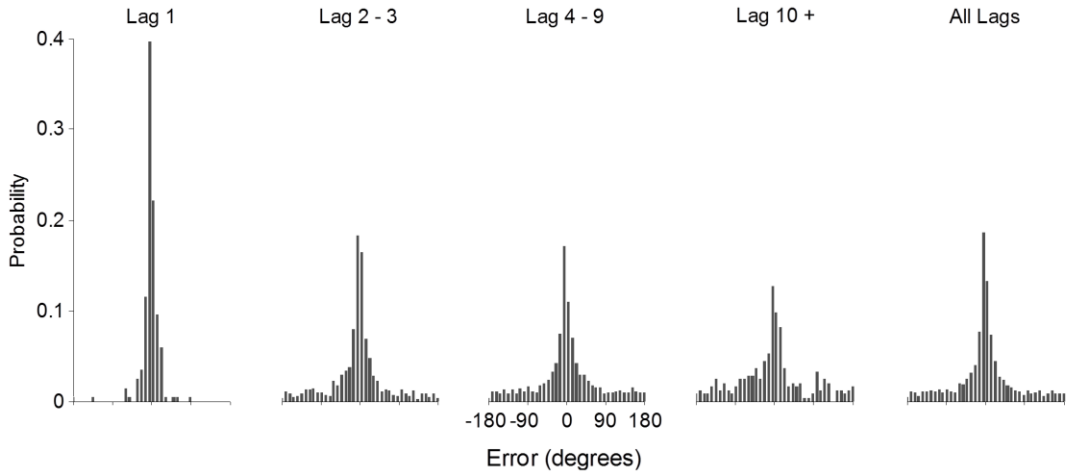

Figure 3.2. Histograms of errors as a function of lag: lag 1, lag 2-3, lag 4-9, lag 10+, as well as the error distribution for all lags.

However, there is also an increased frequency of responses past 50 degrees of error (i.e., increased rate of guessing).

Partitioning the data by lag shows a progression in the decrease of fidelity, and corresponding increase in the rate of guessing, that cannot be discerned from an aggregate error distribution. Furthermore, in the aggregate error distribution (Figure 3.2, 'All lags' panel), the center portion of the error distribution—which under the remember-guess model is characterized by a single Gaussian distribution—appears to have both a sharp peak as well as broad 'shoulders' suggesting multiple components. However, a visual inspection of the error distributions by lag is insufficient to determine whether the composition of the error distribution is strictly that of remembering and guessing, or if there are additional factors at play.

Recall bias To assess bias in recall we calculated the difference between the hue value recalled and the hue value studied. Figure 3.3, top left panel, shows study hue values as a function of bias. The square boxes illustrate the bias for each studied value scaled by the

frequency at which the response was given across participants. Each square box is colored with the true recalled hue value given for each studied value. All responses to a particular studied value form a straight horizontal line, and correctly recalled hue values lie vertically at the zero-bias line of the x-axis. The results shows regression to the mean effects for several color categories, where accuracy is greatest closer to the mean of the categories and hue values greater than the category mean are predominately underestimated (to the left of the zero bias line in Figure 3.3, top middle panel), while hue values less than the mean are overestimated (to the right of the zero baseline in Figure 3.3, top middle panel). Notably, there is an asymmetry in the distribution of responses around the zero bias line within color categories. When there is a large mass of values to the left of the zero bias line (underestimation), there are very few values to the immediate right, and vice versa. This results in strong diagonal bands (tilted on the vertical axis) within categories that are not merely a result of how the data are plotted. We take the asymmetry to indicate regression to distinct categories, and evidence of an influence of prior category knowledge on memory (Hemmer, Tauber, & Steyvers, 2015; Hemmer & Steyvers 2009a; Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, & Vevea 2000; Hemmer, Persaud, Kidd, & Piantadosi, 2015).

Regression analysis Based on established universal categories (red, orange, yellow, green, blue, purple and pink; Berlin & Kay, 1969), we assume that the observed recall bias is toward these seven categories (also, see Persaud & Hemmer, 2014). A linear regression model was fitted to each subject for each category (Figure 3.3, top right panel). Because the regression effect is assumed to operate on memory (not guessing), the data were trimmed to remove responses assumed to be guessing. It is unclear prior to

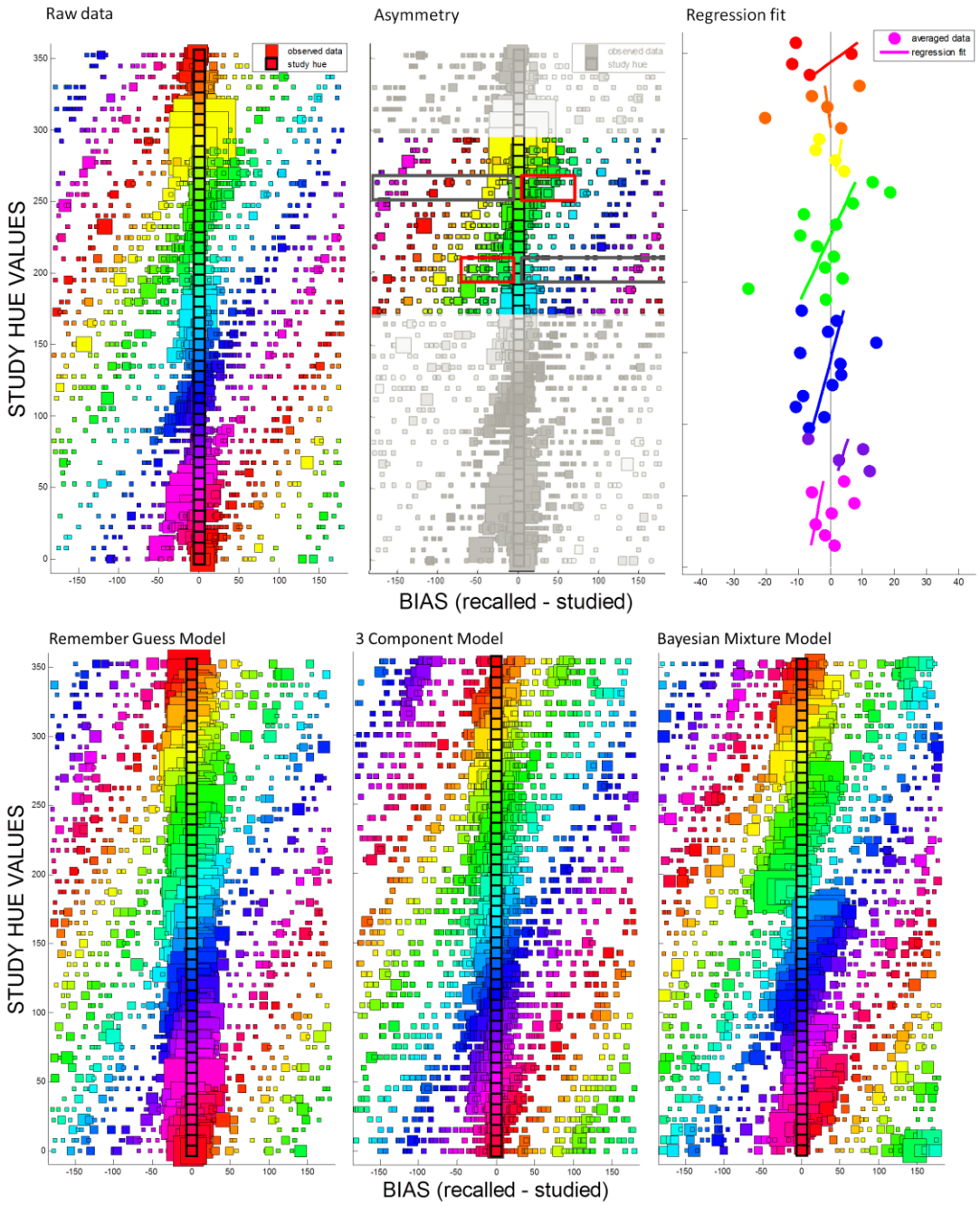

Figure 3.3. Recall response bias to studied hue values. Top left panel: All responses for a given study hue value appear in a horizontal row. The response markers are scaled by the frequency at which they were given (larger boxes indicate greater frequency) and colored with the exact hue value chosen. Top middle panel:

model fitting how to determine a guessing trial, therefore, data were trimmed following

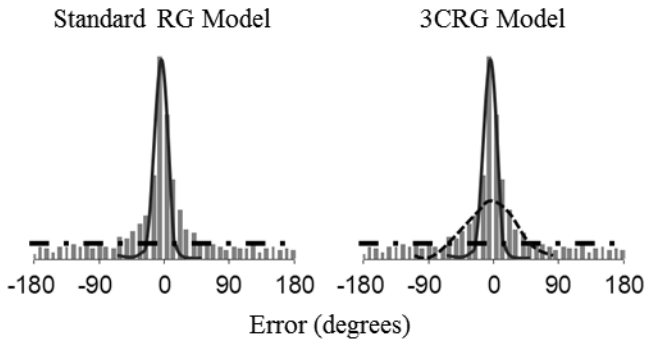

Figure 3.4. Model predictions. Left panel shows stylized predictions of the fit of the standard RG model to the aggregate data. The solid black line represents the memory contribution and the -. dashed line represents guessing. Right panel shows predictions of the fit of a 3CRG model to capture the 'shoulder' of the error distribution. The --dashed black line represents the contribution of the additional

two different procedures: First, singletons in the data, grouped over response frequency, (Figure 3.3, top left panel) were removed and only responses within 75 hue values of the study value were considered for the analysis. A t-test of the subject slopes for each category found that slopes were significantly different from zero for all categories except orange, yellow and purple [red: $t(60)= -2.82$, $p<0.001$; orange $t(60)= -0.15$, $p=0.44$; yellow: $t(60)= 1.96$, $p=0.97$; green: $t(60)= -3.82$, $p< 0.001$; blue: $t(60)= -1.78$, $p= 0.04$; purple: $t(60)=0.94$, $p= 0.83$; pink: $t(60)=-1.65$, $p< 0.05$]. Mean slopes were red: -0.69; orange: -0.02; yellow: 0.27; green: -0.47; blue: -0.17; purple: 0.2; pink: -0.16. Second, guessing responses were trimmed based on the inferred parameters from the Bayesian mixture model to all data. Thus, only responses within 3 standard deviations [$\tau= 25.88$] of the study value were considered for the analysis. A t-test of the subject slopes for each category found that slopes were significantly different from zero for green [t(60)=-5.09, p= 0.00] and blue [t(60)=-1.87, p= 0.03], marginally significant for red [t(60)=-1.38, p= 0.09], but not for orange, yellow, purple or pink. Mean slopes were red: -0.36; orange: 0.12; yellow: 0.54; green: -0.30; blue: -0.10; purple: 0.93; pink: -0.16.

Visual inspection of Figure 3.3 suggests that the lack of significant regression in the purple category might be due to purple not being as salient as any of the other categories. Furthermore, the lack of significant regression in the yellow category might be due to yellow being the smallest category, and because of very high accuracy. Both orange and yellow also appear to have large overlap with the red category.

Modeling

To investigate the components of the error distributions and the observed regression patterns, we implemented two extensions of the standard remember-guess (RG) model: A three component remember-guess (3CRG) model, and a Bayesian mixture (BM) model. The standard RG model assumes that the error distributions are composed of two elements, a Gaussian-like memory distribution and a uniform distribution. See Figure 3.4, left panel for a graphical illustration of predictions for the RG model. Based on our observed pattern of data in the center portion of the error distribution, we predict that the combination of these two distributions will miss some of the area in the error distribution. If the memory component captures the peak of the error distribution, then it may miss the 'shoulders' and vice versa. This combination of a peak and shoulders might signal multiple components in memory.

To explain this pattern in the data, we first implement a simple extension assuming that memory is drawn from two normal distributions, one with high precision and one with lower precision. This additional parameter will allow the model to capture both the peak and the shoulder of the error distribution. See Figure 3.4, right panel for a graphical illustration of predictions for the 3CRG model.

Another pattern in the data that cannot be explained by the RG model, or by our first model extension, is the regression to the mean effect (we will return to a detailed regression analysis in the model comparison sections. See 4.3 Comparison of Regression). The 3CRG model also does not give a theoretical explanation of what the extra component (the low precision memory distribution) represents. Therefore, we implement a second extension—a Bayesian mixture model—which assumes that the additional component is the contribution of prior category knowledge. Rather than the mixture of two separate Gaussian distributions in the 3CRG model, the BM model assumes a single Gaussian distribution where the mean is a weighted linear combination of memory and prior knowledge. This model inherently predicts the regression to the mean effect. However, this effect is obscured in the error distributions, and necessitates the evaluation of the model to the full range of responses, rather than aggregate errors. Next, we detail the implementation of the three models and describe the results of the model comparisons.

Standard 'Remember-Guess' Model (RG)

We implemented the standard RG model using the MemToolbox (Suchow, Brady, Fougnie, & Alvarez, 2013; memtoolbox.org). In this model, the probability density function is given by,

$$(1 - g) * \text{von Mises}(0, \sigma) + g * \text{Unif}(-180, 180) \quad \text{(Eq 1)}$$

Table 3.1. *Model Parameter Values (confidence intervals)*

Remember-Guess Model (RG)

	Fidelity (Conf. Int.) σ (°)	**Guess Rate** (Conf. Int.) *g*
Lag 1	**11.85** (11.85-13.49)	**0.06** (0.03-.010)
Lag 2-3	**16.15** (14.40-18.25)	**0.42** (0.37-0.46)
Lag 4-9	**17.63** (15.82-19.70)	**0.49** (0.46-0.53)
Lag 10+	**15.02** (12.03-20.67)	**0.61** (0.53-0.69)
All	**15.82** (15.03-17.13)	**0.46** (0.44-0.48)

3 Component Remember-Guess Model (3CRG)

	Fidelity (Con. Int.) **σ (°)**	**Fidelity** (Con. Int.) **τ (°)**	**Guess Rate** (Con. Int.) ***g***	**Mixing Parameter** ***w****
Lag 1	**11.44** (9.87-15.84)	**28.76** (13.68-58.43)	**0.03** (0.01-0.09)	**0.28**
Lag 2-3	**15.27** (13.68-17.69)	**27.87** (20.90-36.96)	**0.40** (0.36-0.45)	**0.35**
Lag 4-9	**17.13** (15.12-19.69)	**29.69** (24.12-48.72)	**0.48** (0.44-0.51)	**0.37**
Lag 10+	**15.21** (12.44-32.11)	**35.21** (13.36-73.98)	**0.59** (0.51-0.68)	**0.30**
All	**15.40** (15.24-16.74)	**28.51** (22.95-32.50)	**0.44** (0.42-0.47)	**0.35**

Bayesian Mixture Model (BM)

	Fidelity (Con. Int.) **ψ (°)**	**Fidelity** (Con. Int.) **τ (°)**	**Guess Rate** (Con. Int.) ***g***	**Mixing Parameter** ***w****
Lag 1	**22.12** (17.22-23.85)	**23.43** (14.10-26.00)	**0.05** (0.01-0.09)	**0.49**
Lag 2-3	**19.80** (14.17-19.80)	**21.12** (17.40-22.09)	**0.43** (0.39-0.58)	**0.48**
Lag 4-9	**18.55** (16.11-18.55)	**25.27** (23.81-25.89)	**0.54** (0.54-0.62)	**0.42**
Lag 10+	**20.29** (12.96-27.71)	**250.6** (242.0-254.9)	**0.60** (0.51-0.72)	**0.08**
All	**19.03** (18.77-20.70)	**25.88** (23.68-26.79)	**0.47** (0.47-0.52)	**0.42**

*$w = (1/\tau^2)/ [(1/\tau^2) + (1/\psi^2)]$

where remembered responses are von Mises distributed (due to the circular hue space) with a mean of μ and standard deviation σ. Guessing responses are produced with probability g and are uniformly distributed across the stimulus range from -180 to 180 degrees. Furthermore, because the error distribution is centered on zero μ=0, this parameter will not be considered in this implementation.

Table 3.1 gives the inferred parameters and 95% confidence intervals. Figure 3.3, bottom left panel shows the simulated draws from the posterior of the RG model. According to the model fits, there is a substantial increase in memory noise (σ)—i.e., decrease in memory fidelity—between lag 1 and lags 2-3. Thereafter, memory fidelity appears relatively constant (overlap in confidence intervals between lag groupings). In addition, there is a steady increase in the guessing rate (g) from lag 1 and forward. The model appears to capture the general trend in the data, with the exception of missing the peak of the distribution at some lags and a small portion of the shoulder at others.

Three component 'Remember-Guess' Model (3CRG)

Next, we implement the first extension. We assume that the memory component is itself a mixture of two Gaussian distributions. This is very similar to the Donkin et al. (2014) model, which assumes two components in guessing, where the extra component only applies at retrieval. Our model, in contrast, makes the assumption that the increased noise is attached to the memory component rather than the guessing component. Our memory mixture is also not conditioned on labeling, but rather applies to all trials. While mathematically the two models are equivalent, they differ in the conceptual underpinnings.

In the 3CRG model, we first assume the additional component is related to the memory component in anticipation of the BM model. Second, we use the noise from the two Gaussian components to determine the mixing, rather than assume an additional free parameter. Third, the assumption that the additional component attaches to memory is agnostic about whether the influence of the additional component happens at encoding or retrieval.

The probability density function of the 3CRG model is then given by,

$$(1-g)*\big((1-w)*\text{von Mises}(0,\sigma_{\text{mem}})+w*\text{von Mises}(0,\tau)\big)+ \quad \text{Eq (2)}$$

$$g*\text{Unif}(-180,180)$$

We assume that the mixture of von Mises distributions *w* is based on the fidelity of these two distributions. This is strongly motivated by the assumption of the BM model that the linear weighting is a Bayesian integration such that $w= (1/\tau^2)/ [(1/\tau^2) + (1/\psi^2)]$. For clarity this can be rewritten as $w= \psi^2/ [\tau^2+ \psi^2]$. Using the noise parameters in this way ensures a tradeoff between the two memory components such that when one has high precision it carries more weight, which seems a reasonable assumption of memory. Furthermore, the noise in one of the von Mises distributions is dependent on the noise in the second distribution, $\sigma_{\text{mem}}=\sqrt{(1/[(1/\tau^2) + (1/\psi^2)]}$. Using the noise parameters in this way establishes a difference in the precision on the two von Mises distributions, such that σ_{mem} is always smaller than τ, and that when ψ and τ are the same, the noise on one von Mises is smaller than the other[3].

It should be noted that for completeness, we also implemented a number of other variations of the 3CRG model including versions where: 1) the weighting *w* is inferred but σ_{mem} is still calculated from ψ *and* τ as above, 2) the weighting *w* is calculated as above, but σ_{mem} as the noise on one von Mises is replaced with ψ, treating the noise parameters as independent (See Appendix A4), and 3) all parameters are inferred, that is the weighting *w* is inferred, ψ is the memory noise on one von Mises and τ as noise on the other—that is, σ_{mem} is not calculated as above (See Appendix A5). Alternate version 1) proved to be very unstable at lag 10+ in the hierarchical fitting (see section 4.2 for

[3] As a toy example, if $\psi=20$ and τ=50, σ_{mem}=18.6 If $\psi=20$ and τ=30, σ_{mem}=16.66. If $\psi=20$ and τ=20, σ_{mem}=14.1

hierarchical fitting of the RG, 3CRG and BM models), and we ultimately abandoned this model. Alternate version 2) provided identical patterns in parameter values as the version implemented in Eq. 2 above, identical values of AIC and BIC, and an even better DIC score. However, the consequence of not tying the noise parameter via σ_{mem} means the model is agnostic about which component is the primary and which is the secondary, and for lag 10+ the hierarchical fitting would sometimes switch whether ψ had the smaller value or τ had the smaller value. Alternate version 2) was also unstable in the hierarchical fitting, but adds some interesting insights (See Appendix A5 for discussion). It should also be noted that the 3CRG model (in any of these versions) is very stable when fitted at individual lags, indicating the robustness of the model.

Table 3.1 gives the inferred parameters of the 3CRG model. Figure 3.3, bottom middle panel shows the simulated draws from the posterior of the 3CRG model. Similar to the model fits of the standard RG model, there is an increase in memory noise (σ)—i.e., decrease in fidelity—between lag 1 and lags 2-3, and memory noise stabilizes across remaining lags. The second memory noise parameter (τ) follows a similar trajectory. There is also an increase in the guessing rate (g) from lag 1 and forward. Overall, the 3CRG model posits a similar noise in memory for the σ parameter, and a fairly similar guessing rate, relative to the standard RG model. In this respect, our findings are remarkably consistent with Zhang and Luck (2008), Brady et al. (2013), and Donkin et al. (2014). The failure of the standard RG model to capture the shoulder of the central error distribution is accounted for by the additional noise parameter of the 3CRG model, while simultaneously providing a better fit to the peak of the distributions (for model comparison see Table 3.2).

Bayesian Mixture Model (BM)

Motivated by the experimental results, which show a regression to the mean (see Figure 3.3, top right panel) for a number of color categories, we sought to develop a model that could take into account this behavior. We propose a Bayesian mixture model where recall is a combination of three inputs: noisy representations stored in memory, prior expectations (category knowledge), and random guessing. This approach combines the likelihood from the Bayesian Cognitive model (BCM) developed by Hemmer and Steyvers (2009b) and Hemmer, Tauber, and Steyvers (2015) with the standard RG model of Zhang and Luck (2008). Importantly, in order to visualize the full range of samples from the posterior to demonstrate the regression to the mean effect, we now fit the model to the observed responses, rather than the error distributions.[4]

We extend the RG model by assuming that responses are based on a combination of samples drawn from memory, with probability *w*, and prior expectations, and otherwise, with probability *g*, responses are assumed to be guesses. In the BM model, standard Bayesian techniques (Gelman et al., 2003) can be used to compute the mean of the posterior distribution:

$$Recall \sim \mathrm{N}((1-w) * \bar{y} + w * \mu, \sigma_{\mathrm{mem}}) \qquad \text{Eq (3)}$$

where recall is a weighted linear combination, of samples *y* drawn from memory with noise ψ and some prior expectation with mean μ and standard deviation τ, for the stimulus feature, and with fidelity $\sigma_{mem}=\sqrt{(1/[(1/\tau^2) + (1/\psi^2)]}$. The μ for each category was specified based on the assessment of expectations for color categories in Persaud and Hemmer (2014; See Persaud & Hemmer, 2014 for predictions from the Bayesian model

[4] We also refitted the standard remember-guess model to the full response distribution. See Appendix Table A3. There is no difference in the parameters of this model between the two fittings.

over the true color space).[5] The weights are a combination of the noise in memory and the fidelity of the prior, such that $w = (1/\tau^2)/\ [(1/\tau^2) + (1/\psi^2)]$. The probability density of recall is given by

$$(1-g) * \text{von Mises}\big((1-w)\bar{y} + w\mu, \sigma_{\text{mem}}\big) + \qquad \text{Eq (4)}$$

$$g * \text{Unif}(0,360)$$

Table 3.1 gives the inferred model parameters for the BM model. Different from the model fits of both the standard RG model and the 3CRG model, there is no change in memory noise (ψ) between lag 1 and lags 2-3, rather memory noise is stable across all lags. The noise on the prior (τ) grows slightly from lag 2-3 to lag 4-9 and then jumps dramatically for lag 10+. The weighting w is steady and evenly split between the memory trace and the prior until lag 10+ where, in response to the large increase in τ, it decreases. As in the RG and 3CRG models the guessing rate (g) increases gradually from lag 1 and forward.

Figure 3.3, bottom right panel shows the simulated draws from the posterior of the BM model. Both the RG model and 3CRG model simulations (Figure 3.3, bottom left and middle panels) show a mass of responses near the center zero-bias line and a uniform spread of remaining responses to either side. That is, responses are equally likely to be over and under-estimated regardless of the study hue value relative to the mean of the color categories. See section 4.3 for regression fit to the RG and 3CRG models. Unlike the RG model and 3CRG model, the BM model can capture the regression to the mean effect, where simulated responses for hue values greater than the category means are

[5] Category means: 1.65°, 30.35°, 60.08°, 119.69°, 230.30°, 284.12°, 323.40°

more likely to be underestimated, while values less than the mean are more likely to be overestimated, creating an asymmetry similar to the raw data.

Model Comparison

Comparison by lag

Model comparison between the RG and the 3CRG models was conducted using the *MemToolBox* (Suchow, et al, 2013; memtoolbox.org). The AIC and BIC values for the two models are reported in Table 3.2 (bold font indicates better fits with a difference score greater than 5, while italicized font indicates marginally better fits with a difference score less than 5). Due to the fact that each participant only performed 48 trials with varying lags leading to a sparsity of data for some lags, individual differences were not assessed. The data was pooled across subjects, and subjects were treated as fixed in both AIC and BIC. An improved fit was observed for the 3CRG model for the aggregate error distribution and all lag groupings, except 10+. It seems reasonable that the model comparison favors the standard RG model at lag 10+ given the increase in the guessing

Table 3.2. *AIC and BIC Model Comparisons by Lag Group*

	AIC			BIC		
	RG	**3CRG**	**BM**	**RG**	**3CRG**	**BM**
Lag 1	1663.16	**1657.47**	1915.22	1673.42	*1672.87*	1930.62
Lag 2-3	9264.21	**9249.76**	9647.52	9277.43	**9269.59**	9667.35
Lag 4-9	17090.03	**17074.14**	17477.20	17104.41	**17095.72**	17498.77
Lag 10+	*3182.68*	3182.96	3232.12	**3193.45**	3199.43	3248.58
All	31374.58	**31332.22**	32275.59	31390.22	**31355.67**	32298.79

*Bold font indicates better fits with a difference score greater than 5, while italicized font indicates marginally better fits with a difference score less than 5.

component of the error distribution. The uniform distribution has lifted and could potentially account for the portion of the error distribution that would be accounted for by the second fidelity component parameter. While the improvement is marginal for lag 1 (less than a 5 point difference in AIC between RG and 3CRG), the improvement is substantial for lags 2-3 and 4-9. Memory at lag 1 appears to have a high level of precision and a majority of the performance can be attributed to remembering with little influence of guessing. In contrast, memory is both precise and less precise at other lags and there is a greater rate of guessing. This is consistent with the 3CRG assumption that there is both a memory component with high fidelity and a component with greater noise. The 3CRG model makes it clear that there is additional information in the error distribution that cannot be solely explained by remembering with noise and guessing (i.e. the RG model).

The AIC and BIC values for the BM model are also reported in Table 3.2. No improvement in fit was observed for the BM model relative to either the RG or the 3CRG models. There are several reasons why the BM model might lose out in the model comparison. For example, we assume only one value for tau for all categories, and we specify the color categories based on universal color categories. Furthermore, the weak regression effects in the data allow the 3CRG model to successfully fit all the data without accounting for the regression effects. We discuss all of these reasons along with possible remedies in the discussion section. It is important to note that making allowances for an influence of category information in the BM model produced the characteristic regression to the mean effect which cannot be captured by the two other models, and we still see this as a substantial strength of the BM model. Restricting analysis to error models—while producing an improved fit—leads to very different conclusions about

memory. The regression effect makes it clear that category knowledge plays an important role in recall, and that this must be considered in models of LT memory.

Hierarchical model comparison

Thus far, we have evaluated the models based on fits at the individual lags. It is reasonable, however, to assume that the same model applies to all lags. Therefore, in addition to fitting the separate models for each lag, we also fitted a single hierarchical model to all lag groupings together, for each of the three models. This model treats each of the lag groupings parameters as samples from a normally-distributed population and then infers both best fitting parameters for each lag grouping, as well as the population mean parameter.

Because AIC and BIC are not appropriate for assessing hierarchical models, here we report DIC scores (Deviance Information Criterion; Spiegelhalter et al., 2002, van der Linde, 2005). The DIC is a generalization of the AIC for hierarchical models, which penalizes both for quality of fit and number of parameters. As before, the fitting was conducted using the *MemToolBox.*

The parameters for each of the three hierarchical implementations were essentially identical to the parameters reported in Table 3.1 across all lag groupings. However, due to the sparsity in the data at lag 10+ some of the models are very sensitive to the choice of prior distribution. This particularly affects the BM model in the hierarchical implementation. The DIC for the models were as follows: RG = 31280, 3CRG = 31230, and BM = 32173. This replicates the pattern of model comparison when lags are

estimated separately. There is an improved fit observed for the 3CRG model over both the RG and BM models.

Regression Comparison

To further understand how the models capture the observed data, a regression analysis was performed on the simulations from each of the three models, similar to the regression analysis performed on the subject data (see section *2.2.3*). We simulated draws from the RG model assuming 61 subjects and 48 study hue values as in the experiment (Figure 3.3, bottom left panel). Because the regression effect is assumed to operate on memory (not guessing), only responses assumed to be drawn from memory (within 3 standard deviations [σ= 15.82] of the study value) were considered for the analysis. A linear regression model was fitted to each simulated subject for each of seven universal color categories: red, orange, yellow, green, blue, purple and pink (similar to the regression analysis for the raw data). Recall that, t-tests of the *observed subject* data revealed that the slopes of 4 of the 7 categories were significantly different from zero. In stark contrast to the subject data, one sample t-tests of the RG model slopes failed to find a significant difference from zero in any category. The mean slopes for all categories were: red: -0.04; orange: -0.16; yellow: -0.31; green: 0.09; blue: 0.11; purple: 0.07; pink: -0.05.

We simulated draws from the 3CRG model following the same procedure as for the RG model. Responses within 3 standard deviations [τ= 28.51] of the study value were analyzed. One-sample t-tests of the 3CRG model slopes failed to find a significant difference from zero in all categories, except purple [$t(60)$ =-1.90, p= 0.03]. Note that the observed subject slopes are not significantly different from zero for the purple category, and thus, the 3CRG model does not mirror the subject data for this category. Mean slopes

were red: 0.79; orange: -0.62; yellow: -0.29; green:-0.07; blue: -0.04; purple: -0.98; pink: -0.29.

Lastly, we simulated draws from the BM model following the same procedure as for the other models. Responses within 3 standard deviations [τ= 25.88] of the study value were analyzed. One sample t-tests for the simulated BM model data revealed a similar pattern to the subject data, in that 4 of the category slopes were different from zero (yellow: $t(60)$= -4.26, p=.00; green: $t(60)$= -3.62, p=.00; blue: $t(60)$= -4.18, p=.00; pink: $t(60)$= -3.68, p=.00). The mean slopes for all categories were red: -0.48; orange: -1.07; yellow: -0.17; green:-0.36; blue: -0.37; purple: -0.30; pink: -0.65.

For completeness, we then compared the slopes from the *subject* data for each category to the slopes of the simulated data. We sought to evaluate whether observed regression patterns in the subject data were observed in the model simulations – i.e., in the categories in the subject data where the slopes were significantly different from zero, the model simulations also resulted in non-zero slopes of the same degree. For the RG model, there were significant and marginal differences in slopes, when compared to the subjective slopes, for four categories (red: $t(120)$= -1.74, p= 0.08; yellow: $t(120)$= 2.12, p= 0.04; green: $t(120)$= -2.87, p= 0.00; blue: $t(120)$=-2.38, p= 0.02). This was due to the RG model either failing to predict a regression (red and green), or predicting an effect in the opposite direction of the observed data (yellow and blue). In the three remaining categories, the failure to find significant differences between the model and the subject data was due to the RG model predicting no regression when there was no regression effect in the observed data (orange and purple), or when the regression effect in the data

was weak (pink). In total, the RG model only correctly predicted two categories—yellow and purple.

Similar to the RG model, for the 3CRG model there were significant and marginal differences in slopes, when compared to the subjective slopes, for four categories (red: $t(120)= -2.24$, $p= 0.03$; yellow: $t(120)= 1.72$, $p= 0.09$; green: $t(120)= -2.53$, $p= 0.01$; purple: $t(120)=2.11$, $p= 0.04$). This was due to the 3CRG model either failing to predict a regression (green), predicting the effect in the opposite direction (red and yellow), or predicting a strong regression when there was no observed regression in the data (purple). In the remaining three categories, the failure to find significant differences was due to the 3CRG model predicting a weak, but non-significant regression effect, when there was a weak, but significant effect in the data (blue and pink), or predicting a weak, but marginally significant regression, when there was no regression effect in the data (orange). In total, the 3CRG model only correctly predicted the regression pattern in the blue and pink categories.

In contrast, for the BM model, there was no significant difference in five of the seven categories. This means that the BM model either predicted a regression to the category mean (red, green, and blue) or no regression (yellow and purple) for the same categories as was observed in the data. In one category (pink: $t(120)=2.47$, $p= 0.02$), the observed difference is due to the BM model over-predicting the steepness of the regression, rather than failing to predict the regression effect. Only in one category (orange: $t(120)=3.59$, $p= 0.00$), does the BM model fail to predict the pattern in the subjective data—by predicting a negative slope when the slope in the subject data, although negative, was not significantly different from zero.

In summation, the collective results of the regression analyses suggests that the slopes generated from the BM model more closely resemble the regression behavior in the subject data, compared to both the RG and 3CRG models.

Discussion

Summary We investigated the time course of errors in recall in an effort to understand the components that contribute to LT episodic memory. We employed a novel experimental paradigm and conducted a lag analysis to characterize the influence of category knowledge, and memory over time. We then implemented three distinct cognitive models to evaluate the potential contributing components to memory. Furthermore, we found that there are two important factors in LTM that cannot be accounted for by the standard RG model. In the aggregate, recall reflects a combination of three components: a peaked memory component, a less precise memory component, and a guessing component, capturing the peak and 'shoulders' in the error distributions. In the full response data, recall reflects regression to the mean effects for several color categories, indicating a contribution of prior category knowledge to memory.

The 3CRG model can account for the additional component in memory, and provided a large improvement in the fit over the RG model. The benefit of the 3CRG model is that it has an additional component that can account for a number of mechanisms that might influence LT memory, such as verbal labeling (Donkin et al., 2014), and variable precision in memory (van den Berg et al., 2012). Despite the strengths of the 3CRG model, there is no clear theoretical interpretation of what is encompassed in this component. Moreover, like the RG model, it also cannot capture the regression patterns in the data. The BM model, in contrast, can account for both a second memory

component and the regression patterns, and the BM model also has a theoretical framework for the additional component. It, however, loses dramatically in the model comparisons. The BM model we implemented here is a first pass at understanding the influence of category knowledge, and there are a number of factors that might account for the 3CRG model being favored over the BM model, such as weak regression effects in the data, fragile associations, incorrect category assumptions and other general modeling assumptions. There are also several possible remedies that might improve the BM model and are discussed in the next section. Furthermore, our results have important implications for understanding mechanisms such as decay, sudden death and interference.

Weak regression effects

A key assumption of the BM is the regression to the mean effect. This effect has been demonstrated to be robust in memory (Hemmer & Persaud, 2014; Hemmer & Steyvers, 2009a; Hemmer, Tauber, Steyvers, 2015; Huttenlocher, Hedges, and Duncan, 1991; Huttenlocher, Hedges, & Vevea 2000; Persaud & Hemmer, 2014; Hemmer, Persaud, Kidd, & Piantadosi, 2015). In our data however, using seven universal color categories as a benchmark resulted in poor alignment to the data. The regression analysis revealed that there was no significant regression in three categories, suggesting that the use of universal color categories in the regression assumptions is likely not representative of our data. Furthermore, the fact that the 3CRG model outperforms the BM model, in both AIC and BIC, suggests that the regression effects are weak enough that the inability of the 3CRG model to fit the regression effects is outweighed by its improved fit to the rest of the data. While Persaud and Hemmer (2014) found strong regression effects to all seven universal color categories, they conditioned their regression analysis on responses where

participants also provided the correct verbal label for the study value at test. In other words, they only analyzed data where the participants were able to recall the association between the test cue (shape) and the study color. Here we include all data, which likely includes trials where participants misassociated shape cues to studied colors, guessed, or made some other error. A key test for the flexibility of 3CRG model without accounting for regression effects would be if the model still outperformed the BM model for the finding of differential bias to two separate categories for stimuli studied at the same size (i.e., a large strawberry and a small apple –See Hemmer & Steyvers, 2009a).

We acknowledge that our findings are likely data dependent. There are several possible considerations that might improve the fit of the BM model, or help to lend further support for the strength and flexibility of the 3CRG model. Since the stimuli were drawn from the true hue space, categories had varying sizes. An example of this can be seen in Figure 3.3, top left panel, where the raw data shows high accuracy (large squares) around the yellow category, because this is a very small category. A possible future extension to the BM model would be one that considers variable precision in the Tau parameter (here we have assumed that there is only one value of tau for all categories). This would be akin to the van den Berg et al. (2012) variable-precision model which assumes variability in the precision with which items are encoded, but with variable precision in the categories. This could remedy the weak regression effects in small categories which obscures the importance of capturing the regression effects in other categories.

Alternative color categories

The samples drawn from the posterior of the BM model (Figure 3.3 bottom right) reveals a misalignment between the color categories used to inform the model, and the actual categories borne out in the data. For example, in the data there appears to be two blue categories–light blue and dark blue. However, the BM model only exhibits regression to one blue category –consistent with universal categories. To better understand what color categories participants might have regressed toward in the response data, we conducted a cluster analysis (see Appendix A6). Interpreting the clusters relative to the standard universal color categories, suggests that observers may be using eight categories—five of which can be interpreted relative to the universal color categories: a category composed of red, orange, and yellow universal color values (visualized in red; Figure A1); another category predominately composed of green values; two separate categories for the hue space encompassing blue values (visualized in light blue and dark blue); one category for purple; and one for pink (although pink may contain red values, given the circular nature of the hue space). Interestingly, there were two uniform clusters that span the entire hue range and fell on the top and bottom edges of the graph. These clusters may potentially correspond to the guessing component, or could relate to the large value for τ at lag 10+ in the BM model. Participants also appear to use color categories at various levels in the color hierarchy. For example, participants appear to use the subordinate categories of light blue and dark blue. On the other hand, for colors in the universal red and orange ranges, they use a superordinate color category for warm colors (i.e. a blended category for red, orange, and yellow).

Another interesting feature of the cluster analysis is the natural prediction of regression to the mean behavior in the data. The inherent regression effect learned from the cluster analysis and the use of color categories with different boundaries, provide important constraints for future considerations of Bayesian modeling of color space. While we think that the BM model provides an important theoretical framework in considering regression effects and category influences, continued failures of the BM model even under improved category assumptions would lend further strength to the 3CRG model.

Fragile associations

Another factor that might impact the performance of the models—particularly in the individual lag fits—is that of fragile associations. Modeling paradigms in visual short-term memory have successfully extended the RG model to incorporate task-based components, such as "misassociation" or "misbinding" parameters (Bays, Catalao, & Husain, 2009; Bays, Wu, & Husain, 2011). There are some hints that there might be fragile associations in our data as well.

At lag 10+, precision in the additional component in both the 3CRG and BM models is low and the rate of guessing is high, favoring the RG model. In fact, lag 10+ is the only lag grouping where the 3CRG model loses. This however, might be a consequence of the experimental design. Following standard procedures in color memory paradigms in visual working memory, we deliberately use an experimental design where we assign colors to random objects (e.g., Brady et al., 2013). An important consequence of this design, in conjunction with long lags, in the study of LTM, is that the object color *pairing* might be what is forgotten. In other words, performance at lag 10+ gives the appearance

of a high rate of guessing, not because of a failure to remember the studied hues, but due to a failure of the shape cue to retrieve the correct hue pairing. A natural task with a stronger cue-target association might result in a substantially different pattern of data – one where the rate of guessing is lower. Recent work by Lew, Pashler, and Vul (2015) proposes an interesting new model of fragile associations in LTM. While this is beyond the scope of this paper, given that we cannot assess fragile associations in the current experimental paradigm, we agree that this is an important future direction. Fragile associations might hamper the BM model more than the other models because the behavior looks like guessing, but it has a strong memory trace, albeit bound to the wrong cue. Therefore, the model has a difficult time assigning the behavior, and the role of prior knowledge appears more diffuse.

Interference vs. Decay

Models of memory have varied in their mechanisms of forgetting. Some models theorize that forgetting occurs as a function of decay of memory traces over time (e.g., Barrouillet, Bernardin, & Camos, 2004; Portrat, Barrouillet, & Camos, 2008), while others attribute forgetting to interference (e.g., Lewandosky, Oberauer, Brown, 2009; Neath & Brown, 2012). Our findings appear to provide support for both forgetting mechanisms. First, our results reveal a decrease in memory fidelity (increased noise in the models' σ parameter) from lag 1 to lag 2-3 in the RG model, but in all three models memory fidelity then remains stable across remaining lag groups. This suggests that the memory trace initially suffers some decay during virtually short-term/working memory,

which supports the decay account (Baddeley & Scott, 1971[6]), but is stable into LT memory. This progression in parameters also suggest that— although we are modeling the lag groupings under the assumption that one model should account for all groupings—there is something different about the data at lag 1— namely very high precision, no second component and virtually no guessing, consistent with short-term/working memory.

While memory noise stays steady across lags, guessing (g) increases across lags for all models. In this respect, our findings are remarkably consistent with Zhang and Luck, 2008, Brady et al., 2013, and Donkin et al., 2014. This has led to the interpretation that there is an upper bound on memory noise in LTM, and that memory suffers a 'sudden death' (Brady et al., 2013). Brady et al., 2013, however, could only make this assertion evaluating the transition from working to LT memory. Our design allows us to understand what happens across lags (time) in LTM. For the BM model, the noise in the prior (τ), exhibits a very different pattern from the RG model: τ is steady on lags 1-3, then increases slightly for lags 4-9, but increases dramatically for lags 10+ (a similar pattern can be seen in the alternative implementation of the 3CRG model (Appendix A5) with all parameters inferred). As a result, the weighting (w) of samples from memory and the prior changes across lags. This can be understood as sampling from different granularities of prior knowledge, consistent with hierarchical influences in LTM (e.g., Hemmer & Steyvers, 2009), and the hierarchical nature of colors (Persaud & Hemmer, 2014). On earlier lags One might use a specific prior (e.g., light red or dark red), but on intermediate lags One might use a prior of 'red', and at later lags, where the noise on the prior is very

[6] Although more recent work suggests that forgetting in short-term memory can also be explained by an interference account of forgetting (see Lewandosky, Oberauer, Brown, 2009 for other interference based views accounting for data traditionally thought to support the trace decay account).

large, One might simply use a prior of warm versus cool—or some similar strategy. This progression in parameters in the BM model is contrary to the idea of sudden death. Taken together, our data suggests that not only is fidelity fixed in LTM, but also, category information plays an important role before One resorts to random guessing. Moreover, there is no decay in LTM and no sudden death.

This leaves interference (Neath & Brown, 2012) as the likely mechanism for increased guessing; especially since the trials in our task are interleaved, and the target-cue bindings (color-shape pairing) are arbitrary in nature. Thus, by lag 10+ it is possible that the memory trace (color) is present, but the association to the cue is difficult to retrieve as a result of studying other target-cue combinations. Such an interference explanation is consistent with a fragile association account of memory (Lew, Pashler, & Vul, 2015), where recall is thought to be a combination of remembered information, misassociated information (incorrectly binding targets to cues), and guessing.

Recent work assessing event-based memory in rhesus monkeys lends further credence to interference being the mechanism of forgetting (Devkar & Wright, 2016). Memory accuracy was found to decrease as a function of proactive interference, such that, previously presented stimuli (as far back as 16 trials) interfered with same/different recognition responses. Also, the influence of proactive interference did not change as a function of presentation time between study and test, and inter-trial time. In other words, longer delays between study and test and between trials, where previously studied information would have decayed, did not hamper interference (again, even when the information was studied 16 trials prior).

Serial dependencies are potentially another source of interference that appears as

guessing. Serial dependences refer to the bias in memory that results from information experienced on previous and present trials. It has been demonstrated in visual perception that memory for one item is influenced by accompanying (even task-irrelevant) information and a running average over previous trials (Huang & Sekuler, 2010). Similarly, the perceptual system is serial dependent in that perception is informed by both prior and present information (Fischer and Whitney, 2014).

While serial dependencies may be present at later lags before participants resort to guessing randomly, they are not the source of interference at earlier lags where category information is still available. Hemmer and Steyvers (2009a) showed that in LT memory, the regression to the mean effect is not a result of sequential dependencies. They demonstrated a differential bias when two items from different object priors (e.g., an apple and a strawberry) were studied at the same size. This is also the case in the data presented here (see Figure 3.3, top row middle panel) where there is a differential bias, for example on the boundary between yellow and green, where neighboring hue value results in regression to opposite categories. Sequential dependencies would result in an equal bias towards either category on the boundary dependent on the previous trial (i.e., if previous trial was green bias would be to green but if previous trial was yellow bias would towards yellow). A critically explicit prediction of the BM model is exactly the differential regression at category boundaries as observed in our data.

Given the design of the paradigm used in this investigation, we draw our conclusions with some caution. It is difficult to disentangle the roles of memory decay and interference as mechanisms of errors and forgetting because we do not control for rehearsal (Lewandosky, Oberauer, Brown, 2009; Portrat, Barrouillet, & Camos, 2008),

and trials are interleaved—which could result in intra-sequence interference (Neath & Brown, 2012)—both factors that are required to discern between decay and interference. The decay versus interference differentiation is further complicated by the idea of equivalence, which suggests that both decay and non-decay models provide strong fits to the same data (Neath & Brown, 2012).

Lastly, a contributing factor to memory fidelity, that is not explored in this work but is noteworthy, is the role of intentional forgetting. When participants are instructed to forget certain information in the study stimuli, this leads to a decrease in the probability that the memory trace is retrievable and a decrease in the overall fidelity of the memory trace (Fawcett, Lawrence, & Taylor, 2016). In this way, memory intentions influence the quantity of information encoded into LTM and the quality of the information. Fawcett, Lawrence, & Taylor (2016) modeled this finding using a hierarchical variable-precision mixture model similar to the standard RG model, with the allowance of variability in encoding similar to van den Berg et al (2012).

Conclusions

The implications of the findings from these three models highlight significant characteristics of LT memory. First, consistent with Donkin et al. (2014), there is a clear intermediate stage in LT memory between precise recall and random guessing. While the difference between our 3CRG model and the Donkin et al. (2014) model is a question of technical assumptions, the difference of these two models to the BM model, however, is one of core assumptions, namely that there is an influence of prior knowledge and a regression to known categories. We further argue that at this intervening step, there is a more generalized influence of expectations beyond verbal labeling. Notably, restricting

analysis to error models that mask intervening steps leads to very different conclusions about memory. Deriving conclusions about memory based solely on error distributions is misleading in that it can obscure critical features of memory, such as the influence of prior category knowledge. Therefore, it is important that future research seeks to move beyond the standard remember-guess paradigm for LT memory, and work to elucidate the role of fragile associations and interference. We believe that we have clearly demonstrated that the 3CRG model is robust and consistently outperforms the other models, and that the BM model explains important patterns in the data.

Appendix

A.1 Label vs No-Label Parameters

Table A1. *Parameter estimates and (confidence intervals) for label vs. no-label conditions*

	Label First		Label Last		No Label	
	Fidelity σ (°)	Guess g	Fidelity σ (°)	Guess g	Fidelity σ (°)	Guess g
Lag 1	**12.19** (11.09-13.97)	**0.04** (0.01-0.08)	**9.7** (7.93-14.36)	**0.08** (0.02-0.20)	–	–
Lag 2-3	**13.17** (10.99-17.89)	**0.33** (0.23-0.42)	**8.9** (6.8-18.32)	**0.4** (0.21-0.55)	**17.46** (15.23-20.49)	**0.45** (0.39-0.51)
Lag 4-9	**15.24** (12.56-19.49)	**0.46** (0.38-0.53)	**10.9** (8.88-14.17)	**0.4** (0.31-0.54)	**19.96** (18.10-22.79)	**0.45** (0.452-0.533)
Lag 10+	**15.33** (12.56-19.49)	**0.46** (0.38-0.53)	**29.84** (14.69-93.48)	**0.61** (0.02-0.83)	–	–
All	**13.65** (12.40-15.36)	**0.40** (0.36-0.44)	**10.41** (8.64-12.97)	**0.43** (0.35-0.50)	**18.61** (16.94-20.79)	**0.48** (0.447-0.515)

Table A1 gives the parameter values for the three experimental conditions: Label First (recall color label before generating the color), Label Last (generate a color before recalling color label), and No Label (never provided a color label). For some lag groups, the model had a difficult time converging given the sparsity of the data. Also, there was no lag of 1 in the No label condition.

A.2 Parameters at All Lags

To develop reasonable lag groups we first infer parameter values for all lags using the RG model. Table 3B provides RG model parameter values for each individual lag.

A.3 RG model 0-360

For comparison to the Bayesian Model on the full range of hue values, we implemented the RG model on the same response data space (0-360 degree). There is no difference in the parameters between the two fittings of the RG model. See Table A3 for the inferred parameters.

Table A2. *Parameters for each lag*

	Fidelity σ (°)	**Guess** *g*
Lag 1	11.85	.06
Lag 2	18.13	.37
Lag 3	14.07	.47
Lag 4	18.13	0.44
Lag 5	18.10	0.47
Lag 6	16.09	0.52
Lag 7	17.14	0.56
Lag 8	18.04	0.49
Lag 9	19.38	0.54
Lag 10	14.78	0.63
Lag 11	10.37	0.48
All	15.84	0.46

Table A3. *RG Model Parameter Values (0-360)*

	Fidelity (Conf. Int.) σ (°)	**Guess Rate** (Conf. Int.) *g*
Lag 1	**11.81** (10.59-13.26)	**0.06** (0.03-.010)
Lag 2-3	**16.08** (14.40-18.15)	**0.42** (0.37-0.46)
Lag 4-9	**17.68** (16.05-19.40)	**0.49** (0.46-0.53)
Lag 10+	**15.13** (12.03-20.65)	**0.61** (0.53-0.69)
All	**15.83** (14.82-16.95)	**0.46** (0.44-0.48)

Table A4. *Hierarchical 3CRG Model with independent noise parameters*

3 Component Remember-Guess Model with independent noise parameters						
	Fidelity (Con. Int.) **σ (°)**	**Fidelity** (Con. Int.) **τ (°)**	**Guess** (Con. Int.) ***g***	**Mixing** ***w****	**AIC**	**BIC**
Lag 1	**10.62** (9.41-12.74)	**28.55** (17.05-80.52)	**.03** (.007-.082)	**0.56**	*1657.81*	*1673.20*
Lag 2-3	**13.77** (12.30-16.15)	**29.46** (23.06-44.79)	**.41** (.35-.45)	**0.45**	**9252.46**	**9272.29**
Lag 4-9	**15.30** (13.65-16.83)	**31.26** (23.61-35.52)	**.48** (.45-.52)	**0.44**	**17077.14**	**17098.72**
Lag 10+	**14.02** (12.28-581.89)	**31.06** (11.93-989.33)	**.58** (.53-.69)	**0.45**	3183.15	3199.61
All	**13.86** (13.01-15.07)	**29.60** (25.04-39.40)	**.44** (.41-.47)	**0.48**	**31339.02**	**31362.48**

A.4 3CRG model without independent noise parameters

An alternate version of the 3CRG model where the weighting $w = \psi^2 / [\tau^2 + \psi^2]$, but σ_{mem} as the noise on one von Mises is replaced with ψ, was implemented. Instead of the noise component in one von Mises distribution being dependent on the other, here the noise parameters were treated as independent.

The probability density function of the 3CRG model with all inferred parameters is given by,

$$(1-g) * \big((1-w) * \text{von Mises}(0,\tau) + w * \text{von Mises}(0,\psi)\big) + \quad \text{Eq(A1)}$$

$$g * \text{Unif}(0, 360)$$

where g , ψ and τ are all inferred values from the data. Table A4 shows the parameter values for each lag group under this model and reports the AIC and BIC scores relative to the RG model. The parameter values for this model in the hierarchical fitting detailed in the modeling section (see section 4.2) were identical to the individual lag fitting and are not reported. Note that the parameter values at lag 10+ had a tendency to reverse in different runs of the hierarchical model, such that sometimes $\psi \approx 14$ and $\tau \approx 30$, but at other times $\psi \approx 30$ and $\tau \approx 14$. Irrespective of the order of the parameter values this version

of the 3CRG model consistently had the lowest DIC=31177.53 in the hierarchical

Table A5. *3CRG Model with all parameters inferred, and with AIC and BIC Scores*

3 Component Remember-Guess Model with all parameters inferred					
Fidelity (Con. Int.) **σ (°)**	**Fidelity** (Con. Int.) **τ (°)**	**Guess** (Con. Int.) ***g***	**Mixing** ***w****	**AIC**	**BIC**
6.97 (5.02-11.84)	**19.95** (13.43-47.71)	**0.03** (.008-.082)	**0.56**	*1657.87*	1678.39
8.45 (6.66-11.27)	**28.40** (23.23-46.63)	**0.36** (0.30-0.41)	**0.45**	**9229.60**	**9256.03**
9.57 (7.66-12.45)	**31.05** (24.76-44.41)	**0.43** (0.38-0.48)	**0.44**	**17051.95**	**17080.72**
13.07 (10.37-19.02)	**94.68** (25.88-439.98)	**0.28** (0.03-0.65)	**0.45**	3312.54	3334.49
9.04 (7.71-10.63)	**29.64** (24.96-39.28)	**0.40** (0.36-0.43)	**0.48**	*1657.87*	**31314.45**

implementation of this model, which favors this model relative to both the RG and BM models.

A.5 3CRG model with all parameters inferred

The 3CRG model with all the model parameters inferred was implemented. Instead of the noise component in one von Mises distribution being dependent on the other, here the noise parameters were treated as independent and *w* is an additional free parameter.

The probability density function of the 3CRG model with all inferred parameters is given by,

$$(1-g)*\big((1-w)*\text{von Mises}(0,\tau)+w*\text{von Mises}(0,\psi)\big)+ \quad \text{Eq (A2)}$$

$$g*\text{Unif}(0,360)$$

where g, w, ψ and τ are all inferred values from the data. Table A5 shows the parameter values for each lag group under this model and reports the AIC and BIC scores relative to the RG model. Table A6 shows the parameter values for this model in the hierarchical fitting detailed in the modeling section (see section 4.2). Note that the parameter values at lag 10+ changes between the two implementations of this model. The

Table A6. *Hierarchical 3CRG Model with all parameters inferred*

	Hierarchical 3CRG$_{all}$			
	Fidelity (Con. Int.) **σ (°)**	**Fidelity** (Con. Int.) **τ (°)**	**Guess** (Con. Int.) ***g***	**Mixing** ***w****
Lag 1	**7.06** (6.40-13.61)	**19.50** (6.86-140.95)	**0.04** (0.01-0.12)	**0.56**
Lag 2-3	**8.38** (8.38-14.26)	**28.65** (28.65-149.05)	**0.36** (0.01-0.38)	**0.45**
Lag 4-9	**9.68** (8.58-15.99)	**31.68** (20.68-131.25)	**0.44** (0.29-0.53)	**0.44**
Lag 10+	**13.19** (.25-15.45)	**114.68** (21.95-114.68)	**0.00** (0.00-0.63)	**0.33**

DIC score of the hierarchical version of this model was 31194.54, which still favors this model relative to both the RG and BM models.

A.6 Cluster analysis The cluster analysis was implemented to infer the categories participants regressed to in the experiment. Briefly, the clustering algorithm (Fraley & Raftery, 2006) performs a hierarchical agglomeration to maximize the classification likelihood for up to 9 groups in each model. Next, the Expectation-Maximization (EM) algorithm calculates the maximum likelihood estimation for all models and number of cluster combinations. Lastly, the algorithm computes the BIC scores for each cluster mixture model with optimal parameter values and returns the best fitting cluster size model. The best BIC score (BIC = -66210.13) revealed that 8 clusters produced the most

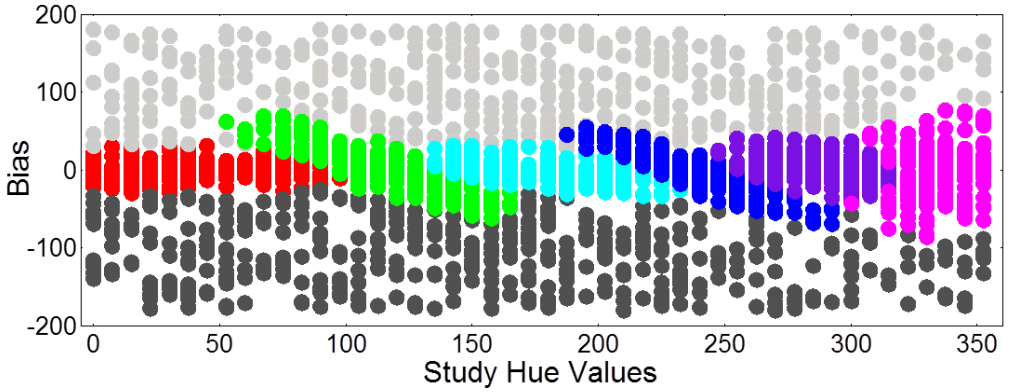

Figure A1. The 8-group unconstrained model-based classification of the data. 8 optimal clusters were learned from the Expectation-Maximization algorithm evaluated by BIC scores. Each of the 8 colors correspond to a different cluster that is color coded to reflect the color category to which most of the study values in the data belong.

optimal partitioning of the data. Figure A1, shows the output from the learned clusters. The mean of the 8 inferred clusters (in degrees) were: 39.59, 151.76, 171.21, 113.88, 182.15, 238.42, 288.69, and 335.57).

Chapter 5: Misassociations, Random Guessing, and Prior Knowledge

K. Persaud took the lead on the study concept and study design with feedback from the advisor, P. Hemmer. K. Persaud performed the stimulus creation, and supervised undergraduates in the lab on testing and data collection. K. Persaud performed the data analysis and the interpretation of the analysis was performed by K. Persaud and P. Hemmer together. K. Persaud wrote the chapter with feedback from the advisor P. Hemmer.

The memory system stores associative information, such as the relationships between information (e.g. the placement of objects to locations in space) over time. The nature of the associative information, such as the semantic coherence and meaningfulness of the associations may impact how this information is stored. On the one hand, previous research suggests that although memory traces can be quickly formed and retained for a long time, memory associations are slowly formed and are quickly forgotten (Lew, Pashler, & Vul, 2015). On the other hand, other work suggests that prior meaningful associative information in the stimulus environment influences recall of current episodic events and improves average accuracy (Hemmer & Steyvers, 2009a; Hemmer, Persaud, Venaglia, & DeAngelis, 2014; Persaud & Hemmer, 2016). These contradictory findings can be reconciled by understanding the nature of the associative information. While arbitrary associations (e.g., the location of objects or color squares in a circle) are fragile in nature, accounting for the difficulty in forming and retaining information, we hypothesize that semantically coherent associations that may have basis in the real world (e.g. airplane in the sky) might produce an opposite effect. When participants bring this knowledge to a task, associations might be learned more quickly

and allow for more strategic recall when retrieving associations from memory. Here, we present the findings from three experiments assessing memory for associations that vary in degree of meaningfulness.

Introduction

In the past decade, there has been an explosion of computational models that characterize the behavior of visual working memory(WM), and have been used to characterize to long-term memory (LTM) (e.g. Bays, Catalao, Husain, 2009; Bays, Wu, & Husain, 2011; Brady, Konkle, Gill, Oliva, & Alvarez, 2013; Donkin, Nosofsky, Gold, & Shiffrin, 2014; Fougnie, Suchow, & Alvarez, 2012; van den Berg, Shin, Chou, George, & Ma, 2012; Zhang & Luck, 2008). These frameworks were developed to explain several memory phenomena, including capacity limits (Alvarez & Cavanagh, 2004; Wilken & Ma, 2004; Zhang & Luck, 2008), memory fidelity (Brady, et al., 2013; Persaud & Hemmer, 2016), the influence of task-dependent factors (Bays, Catalao, Husain, 2009; Bays, Wu, & Husain, 2011; Donkin, Nosofsky, Gold, & Shiffrin, 2014), and the role of pre-experimental category knowledge (Bae, Olkonnen, Allred, & Flombaum, 2015; Persaud & Hemmer, 2014; Persaud & Hemmer, 2016). Importantly, a prominent assumption of many of these models, particularly those that use and extend the popular remember-guess framework (Zhang & Luck, 2008), is that random guessing is a major contributor to recall performance. However, when trying to recall information from memory, do people really guess randomly?

Consider the following example: imagine telling your friend about a scuba diving excursion during your last tropical island vacation. You are trying to recall as much as possible, but the vacation was a long time ago and the details are a bit blurry. Some of your episodic memory representations might be clear and precise (i.e. encoded with high

fidelity), and you recall fish, dolphins, coral, and sea turtles. Other representations might be less clear (i.e. encoded with less fidelity) making them difficult to retrieve from memory, so you simply guess randomly that you saw a coffee cup. *What?* Now, it is not completely unlikely that you saw a coffee cup - one could have fallen off a boat - but it is improbable. It is also unlikely that you would resort to completely randomly guessing[7]. A completely random guess in this case is not constrained to guessing an object, but instead all possible things including concepts such as love or justice (i.e. sampling from an unconstrained distribution over all possible responses in the world). In this example, it seems more likely that you would guess with meaningful information, such as seaweed or sand - things that you would find in the scuba diving context (i.e. sampling from a constrained prior distribution over possible responses in the task, assigning more probability to the objects that are most related to the studied object). While random guessing in real world contexts seems like an unlikely strategy to use, notably many of the aforementioned models assume memory to be strictly a combination of noisy

[7] The term random guessing can be ambiguous and requires both a theoretical and computational description. Theoretically, a random guess is a sample drawn from an unconstrained uniform distribution over all possible features or concepts in the real world. This is analogous to the example given above, i.e. randomly guessing the concept love, peace, or coffee cup. However in a computational model, such as those discussed in this paper, a random guess is a sample drawn from a constrained uniform distribution over all possible response options within a task. For example, a random guess could be to guess from the set of available objects or the colors present on a color wheel. To use prior knowledge is to sample from a prior distribution that might assign more probability to certain response options relative to others. For example, if an individual studies an object above an island, they might guess using only objects that belong above the island (using a prior distribution). If they guess randomly within the computational interpretation of a random guess (i.e. constrained to task based response options), they might guess uniformly over all objects available (regardless of where they belong relative to an island). Importantly, current studies using the remember-guess paradigm (e.g. Brady et al, 2013; Lew et al, 2015; Zhang & Luck, 2008) fail to make this distinction. They discuss random guessing in terms of the theoretical interpretation (unconstrained guessing), but the model assumes the computational interpretation (constrained). It is unclear which interpretation is favored in these studies. In terms of describing behavior, i.e. what people do when trying to recall information from memory, claiming that people guess randomly in the world, but guess in some constrained way in the task is misleading about how memory actually works.

memories and random guessing. If people are not guessing randomly, then what are they doing?

One potential explanation is that people are not guessing randomly, but instead are producing misassociations (Lew, Pashler, & Vul, 2015). Misassociations occur when previously studied non-target information interferes with the recall of target information at test. In this way, individuals can either recall target information with noise, misattribute or miss-bind target and cue information, or guess randomly. In the context of the earlier example, this is equivalent to recalling details, such as seeing a Morey eel, but from a different scuba trip. Lew et al. argued that some information (e.g., location) is easy to store and difficult to forget, but associative information (i.e., memory for the associations of individual items to studied locations) is difficult to store and easy to forget, leading to errors of misassociation.

Another explanation is that people are not guessing randomly, but are using pre-experimental prior knowledge. The use of prior knowledge might be reflected in the finding of a lack of random guessing in a memory task, and instead can reflect informed guessing which assigns more probability to response options that better fit the to-be-remembered information (e.g. guessing sting ray in the context of scuba diving). Prior knowledge has been shown to exert a strong influence on recall, when the to-be remembered information is ecologically valid and meaningful (Hemmer, Persaud, Kidd, & Piantadosi, 2015; Hemmer, Persaud, Venaglia, & DeAngelis, 2014; Hemmer & Steyvers, 2009a; Steyvers & Hemmer, 2012). The influence of prior knowledge has been characterized by Bayesian cognitive models of memory which assume that prior expectations for the statistical regularities of the environment are integrated

with noisy episodic content at recall (Hemmer & Steyvers, 2009b; Hemmer, Tauber, & Steyvers, 2015; Huttenlocher, Hedges & Vevea, 2000; Persaud & Hemmer, 2014; Persaud & Hemmer, 2016; Shiffrin & Steyvers, 1997; Steyvers, Griffiths, & Dennis, 2006). These models, assuming prior knowledge influences memory, suggest that events can be reconstructed based not only on memory, but also on category knowledge and expectations for items naturally associated with the context (in the scuba diving example: fish, coral, etc.). The influence of prior knowledge has been demonstrated in a number of cognitive domains, including word learning (Xu & Tenenbaum, 2007), attention (Kim & Rehder, 2011), human gaze control (Henderson, 2003), functional memory capacity (Ricks & Wiley, 2009), incidental category learning (Clapper, 2012), reading comprehension (Bransford & Johnson, 1972), perceptual categorization (Huttenlocher, Hedges, and Duncan, 1991; Huttenlocher, Hedges, & Vevea 2000; Jern and Kemp, 2013; Galleguillos and Belongie, 2010), among others.

Determining the role of random guessing in LTM and the circumstances under which it may, or may not, happen places constraints on experimental design and the applications of theories of WM to LTM. In this paper, we seek to challenge prevailing theories of visual long-term memory that assume that random guessing is a large contributor to memory performance. We hypothesize that a larger portion of errors in memory result from misassociations and an influence of prior knowledge, *not* random guessing, when stimuli are aligned with people's expectations. When individuals use prior knowledge in tasks, they may guess strategically (i.e. not randomly) when trying to recall information. The ability to use any form of prior knowledge might reduce that amount of random guessing (i.e. assigning uniform probability to all response options within a task).

The goal of this work is to first establish, experimentally whether participants have expectations for the placement of objects to location in space, particularly in an artificial laboratory based stimulus environment and then experimentally assess if the behavior of memory (i.e. whether people recall information correctly, missassociate to other studied information, or guess randomly) differs when people have expectations for the to-be-remembered information (meaningful associations) relative to when they do not(random associations). The direct contribution of the types of responses individuals may make during recall cannot be observed experimentally in this paradigm, and therefore, we employ a model of misassociations (Lew et al., 2015). Since the goal of this paper was to determine if a simple manipulation of the stimuli can result in a change in the behavior of memory, we directly implemented the Misassociations model and its current form outline in Lew et al., 2015. We acknowledge that there may be a host of other models that better describe performance in the task, and discuss limitations of using this model and discussions modifications that could be made.

Here, we quantify the combined contribution of misassociations, prior knowledge, and random guessing to long-term memory. We bring together two lines of research (one evaluating the role of prior knowledge and the other evaluating misassociations) to provide a comprehensive explanation of long-term memory. Understanding the role of prior knowledge relative to misassociations and random guessing has important implications for theory development of factors that contribute to long-term memory and future modeling implementation.

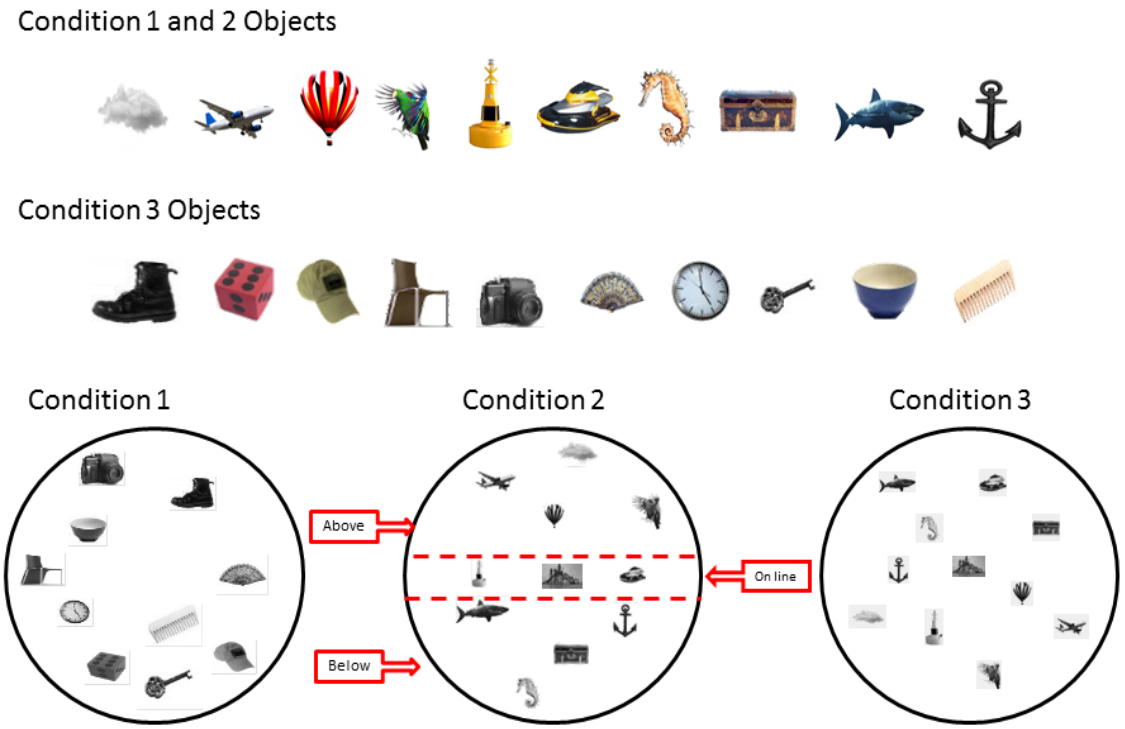

Figure 4.1 Experimental Stimuli. The first two rows show the objects that were used in each condition. The third row shows sample study locations of objects in each condition.

Experiments

To assess the contribution of misassociations, prior knowledge, and random guessing to long-term memory, we employed the experimental design of Lew, Pashler, and Vul (2015) where observers performed cued recall for the locations of 10 discrete objects presented in a circle. Similar to previous work, we started our investigation by first assessing people's prior expectations for the stimulus environment under the assumption that people bring those expectations with them to the task of remembering (Experiment 1). In experiments 2 and 3, we followed the training and testing procedures of Lew et al., 2015. In their original design, they trained participants on the locations of 10 semantically *unrelated* objects presented in random locations around a circle, and

assessed memory for object-location pairings learned during training. In experiment 2, we adopted this training procedure, and in experiment 3 we also tested memory. Both experiment 2 and 3 had three conditions, carried out between subjects. In condition 1, we sought to replicate the findings of Lew et al., by training participants on the 10 *unrelated* objects presented in random locations (see Figure 4.1). In condition 2, we used 10 objects that were semantically *related* to a center object and presented in semantically coherent locations around a center object. This approach allowed us to evaluate the effect that meaningful information had on forming and storing the associations in memory.

In condition 3, we used the same semantically related objects from condition 2, and presented them in random locations. This condition allowed us to disambiguate the contribution of the semantic relationship of the objects from the coherence of the object-location associations on long-term memory. In Experiment 3, we adopted a modified training procedure and tested memory for object-location associations over time. This procedure also mirrored the methodology of Lew et al. We chose to mirror their approach in order to compare performance in our tasks with that of theirs. See Table 4.1 for a breakdown of experimental conditions.

Methods

Participants One-hundred and thirty-six Rutgers University students participated in this study for either monetary compensation or course credit. Twenty-eight students participated in the prior knowledge task for Experiment 1. Forty-nine students participated in Experiment 2 (15 in condition 1, 18 in condition 2, and 16 in condition 3). Fifty-nine students participated in Experiment 3 (19 in condition 1, 20 in condition 2,

and 20 in condition 3). Two participants failed to complete all three sessions of the study (both from condition 2) and were therefore, omitted from the testing analysis.

Stimuli The 10 semantically unrelated objects for condition 1 were selected for the Lew et al., 2015: boot, die, baseball cap, camera, fan, clock, key, bowl, comb, and chair. We selected the 10 objects for conditions 2 and 3 to be semantically related to an island scene, both in terms of meaning and location. The objects were: an airplane, shark, cloud, hot air balloon, buoy, bird, jet-ski, treasure chest, seahorse, and an anchor. The images of the objects were 60 x 60 pixels in size and were presented within a circle with a 50 x 50 pixel island image in the middle, serving as a landmark for the center of the circle. The circle had a radius of 450 pixels. In Experiment 3, conditions 1 and 3, the objects were placed in random locations around the circle. In condition 2, object placement was constrained by the location association to the island: four objects appeared in the section above the center island – which could be interpreted as the sky (airplane, cloud, hot air balloon, bird), 2 appeared on line with the island (buoy, jet-ski), and four appeared in the section below the island – which could be interpreted as the water (shark, treasure chest, seahorse). The display of the circle and objects was maximized across the entire computer screen. The experiment was written in the Matlab computer programming language and all participants performed the study in the laboratory setting.

Design & Procedure

Experiment 1

In the prior knowledge task, participants were presented with two sets of objects, one set composed of the island objects and the other set composed of the mixed items from Lew

Table 4.1 *Experiments and Conditions*

Condition	Experiment 1	Experiment 2	Experiment 3
Condition 1 (Fully Random)	Related and unrelated objects	Training only; unrelated objects in random conditions	Training and testing unrelated objects, random locations
Condition 2 (Fully Associated)		Training only; related objects in coherent locations	Training and testing related objects, coherent locations
Condition 3 (Partially Associated)		Training only; related objects in random locations	Training and testing related objects, random locations

et al. (2015). The sets were presented in random order across participants. Participants were shown a large circle in the center of the computer screen. The circle contained an image of an island directly in the center. Each object was presented one at a time in top left corner of the screen (outside of the circle). Participants were instructed to place the object in a location inside the circle where they would normally expect to find it. After placing all the objects from one set, there was a brief inter-stimulus interval where the screen was blank for 1s, followed by the presentation of the objects from the second set. This task was self-paced.

Experiment 2-3

Experiments 2 and 3 investigated learning and memory for the locations of objects. Participants were given a cover story that they had just returned from a sight-seeing trip on an island, and had convinced their friends to visit. The friends are now visiting the island, and the participants must remember the location of all the exciting objects to point out to their friends. The experiment was carried out in three phases: a training phase, a

distracter task, and an object location recall phase. For the recall phase, participants were tested on the same day of training, one day later, and seven days later.

Training Phase Following the methodology of Lew et al. (2015), Experiment 2 was training only and Experiment 3 was both training and recall. In both Experiments, individual objects appeared outside of the circle, in the upper left corner, and participants were instructed to learn the location of the objects. They responded by clicking within the circle. A red crosshair appeared in the location the participant selected, and the object appeared in its correct location as feedback. For experiment 2, participants followed this procedure for all ten objects for a total of 20 blocks. In other words, each object was practiced 20 times with feedback. For experiment 3, participants only advanced to the next block if all of the objects were placed correctly. An object was classified as correct if the participant responded within 50 pixels of the true location. Correctly located objects were dropped out of future trials in the current block, and incorrectly located objects were repeated as another trial within that current block. A participant advanced to another training block once all object locations were correctly reconstructed in the current block. A participant completed the training phase after 5 blocks. Responses in all training blocks of both Experiments were self-paced.

Distractor Phase (Experiment 3 only) Once the training phase was completed, participants completed a distractor task, consisting of 12 arithmetic problems utilizing two operands (+ or -), and numbers from 0 to 40. Participants were instructed to solve the problems as quickly and accurately as possible. They were not provided with a calculator. The distractor phase was also self-paced.

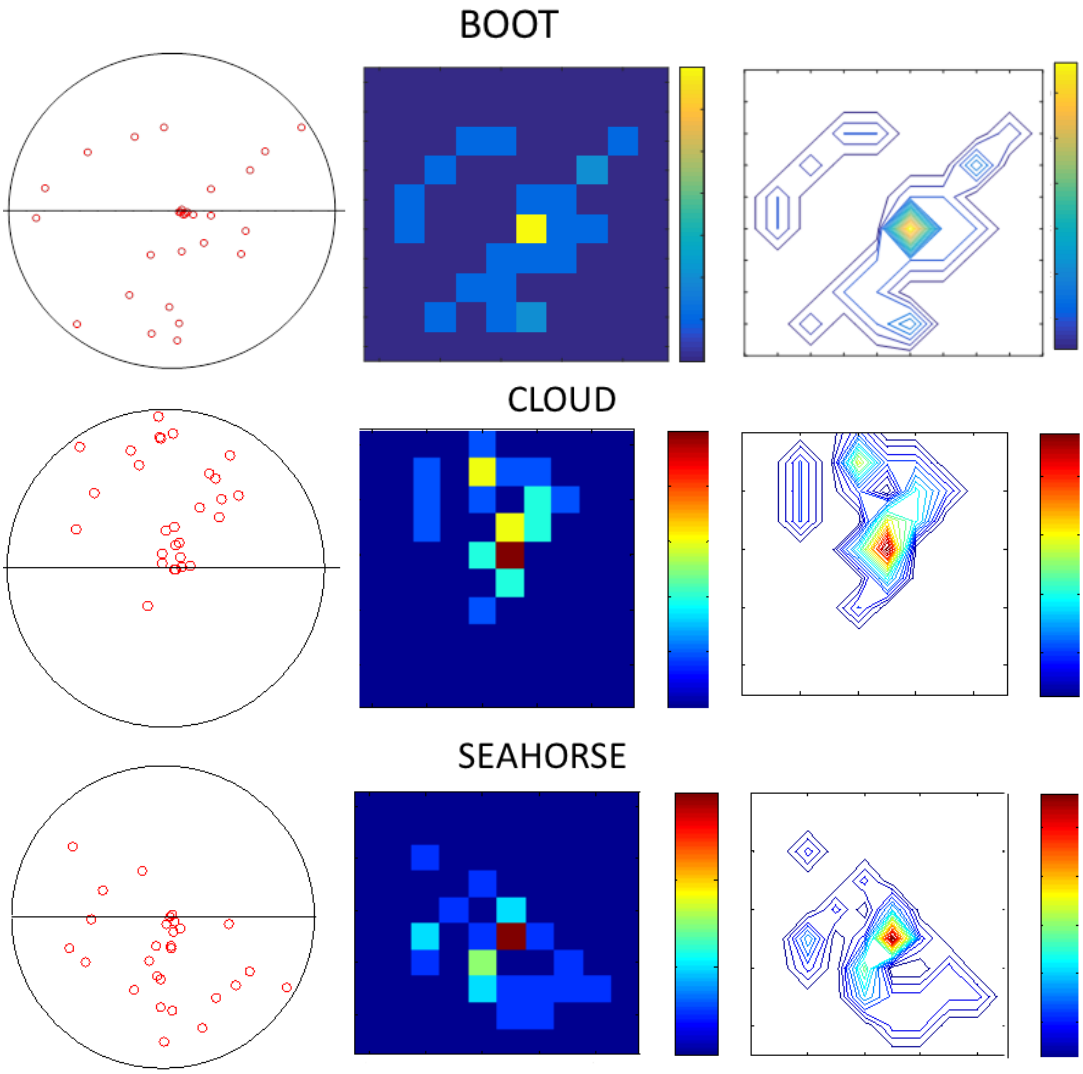

Figure 4.2 Semantically related and unrelated objects. Boot is an unrelated object (condition 1). Cloud is a related objects object found prototypically above the island and the seahorse is a related object found prototypically below the island (conditions 2 and 3). Placement locations, Heat maps, and Contour plots.

Testing Phase (Experiment 3 only) After providing responses for all 12 distractor problems, participants were instructed to recall the studied locations of each object learned during the training phase. Similar to the training phase, participants were shown each object, one at a time, in the upper left corner outside of the circle. The island

was located in the center of the circle serving as a spatial reference. They clicked in the circle where they recalled studying the object during the training phase, and a red crosshair appeared where they had selected. Unlike the training phase, no feedback was provided. This phase consisted of only one block, and was complete after the participant recalled the locations of all 10 objects. For the two subsequent testing days (i.e., one day later and 7 days later), the study mimicked the first day except there was no training phase or distracter phase, and no feedback was given after the participants recalled object locations. The recall task was self-paced.

Results

Experiment 1

We first present data from the prior knowledge condition (experiment 1). Figure 4.2 shows heatmaps and contour plots for the random objects (condition 1) and semantically associated island objects (conditions 2 and 3). The heatmaps and contour plots visually support the idea that people have expectations for where objects belong that are consistent with the coherent locations of these objects in the real world. For example, the airplane and hot air balloon were generally placed in the northern hemisphere of the circle, i.e., above the center island object – which could be interpreted as the sky. Conversely, the seahorse and anchor were generally placed in the southern hemisphere, i.e., below the island object – which could be interpreted as the water. In fact, out of the 112 responses for the above and below objects, only 15% of the objects of either type (above or below) violated the expected placement relative to the island. Above objects were expected to be placed above the island (i.e. having a y-coordinate greater

than 500 which is the y-coordinate for the island) and the below objects were expected to be placed below the island (i.e. having a y-coordinate less than 500).

Given that the semantically related objects were associated with prototypical locations along the y-axis, we tested for a difference in the y-coordinate placement of these objects. There was a significant difference ($t(6)=8.18$, $p<.001$) in the y-coordinates for objects that prototypically belong above the island (i.e., cloud, hot air balloon, airplane, and bird) compared to the objects that are prototypically found in the water below an island (i.e., seahorse, anchor, shark, and treasure chest). Also, a one-sample directional t-tests revealed that the y-coordinates for the above objects were significantly greater than 500, which is the level of the 'horizon line' of the island ($t(111)=9.70$, $p<.001$) and the y-coordinates for the below objects were significantly less than 500 ($t(111)=-6.99$, $p<.001$).

The remaining two objects (i.e., the buoy and jet ski) could be prototypically found on the horizon line of the island, which was located at 500 on the y-axis. A one-sample t-test revealed no difference in the y-coordinates of these two objects from 500 ($p>.1$). Taken together, these results suggest a strong subjective agreement of the prior locations for the semantically related objects. These expectations were elucidated in a relatively arbitrary task (i.e., placing objects in a circle), where little prior instruction was given. Participants clearly have pre-experimental expectations for stimuli that can potentially impact performance in even simple cognitive tasks. It should be noted, however, that this agreement on object placement only applied to placing the objects above or below the island. These expectations are not constrained to how high above or far below the objects were placed relative to the island. Neither did it constrain the placement of objects to the

right or the left of the island (along the x-axis). This resulted in some variability of object placement, even though the objects were placed in the coherent hemispheres. In this way, an influence of expectations borne out in the subsequent memory task might only reflect placing the objects within the coherent hemisphere, and not necessarily at the true location within the hemisphere. See appendix section 1 for visualizations of the memory data.

Unlike the semantically related objects, the semantically unrelated objects for condition 1 did not have a strong prior association for belonging above or below the island. As a result, a majority of the random objects were placed in close proximity to the island (see Figure 4.2). While a one-sample t-test revealed that the y-coordinates for the random objects were significantly different from 500 ($t(6)=8.18$, $p<.001$), the confidence interval suggests that responses were only slightly above the horizon line of the island. Of the 280 responses to the semantically unrelated objects, 50% fell above the island and 49%, fell below the island (the remaining 2% fell directly on the horizon line). This roughly 50/50 split suggests no strong bias to placing the unrelated objects in either hemisphere around the island. The difference in expectations for the two sets of objects foreshadows a difference in memory performance in subsequent tasks.

Experiment 2-3

As a roadmap of the results for experiments 2 and 3, we first compared memory performance across the three testing days within each condition. Next we calculated the memory reaction times across days and conditions. After, we implemented Lew et al. (2015) Misassociation model to learn the contribution of noisy memory, misassociations, and random guessing to performance in each condition. Memory

performance during both training and testing was evaluated by calculating root mean square error (Figure 4.3). The error measure used was the Euclidean distance between studied and recalled object locations[8].

Comparisons of Recall Within Condition

A difference in error across testing days might suggest a difference in the contribution of noisy memory, misassociations, or random guessing over time. A series of analysis of variance (ANOVA) tests were conducted to evaluate root mean square error (RMSE) across the three test days *within* each condition (Figure 4.3). For condition 1, where the objects were all semantically unrelated and presented in random locations (replication), there was a marginally significant difference in error across test days (F[2, 42] = 2.99, p=.06), such that earlier test days had less errors (Day1: M=47.72[36.6]; Day2: M=67.37[64.6], Day3: M=105.55[86.53]). Planned post hoc analyses revealed a significant difference between testing days 0 and 7(t(28) = -2.38, p<.05), but not between days 0 and 1(p>.05) and 1 and 7(p>.05).

Similarly, for condition 2, where the objects were all semantically related and presented in semantically coherent locations, there was a significant difference in error across test days (F [2, 45] =4.59, p<.05), such that earlier test days had less errors (Day1: M=45.14[26.94]; Day2: M=50.26[20.3], Day3: M=76.07[41.66]). Planned post hoc analyses revealed a significant difference between days 0 and 7(t(30) = -

[8] Since the goal of this paper was to evaluate whether the types of error in long-term memory change as a function of the stimulus associations, we chose to evaluate performance using RMSE which allowed us to implement the Misassociations model. However, RMSE does not provide information about the direction of responses given to studied locations (e.g. did participants place objects studied above in the above location or somewhere else in the circle). Therefore, we visualize and discuss object placement during recall. See appendix section 1.

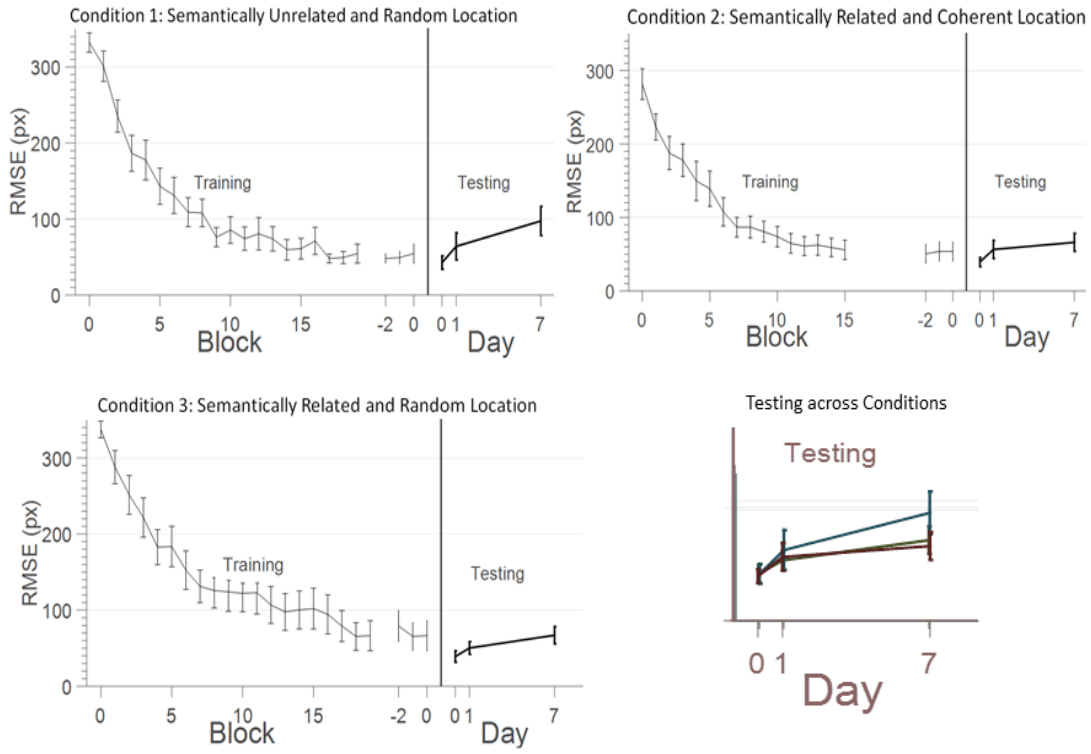

Figure 4.3 Error Root mean square error (RMSE) for training and testing across conditions. Bottom right panel shows testing from the three conditions overlaying each other. Blue is condition 1, red is condition 2, and green is condition 3.

2.49, $p<.05$) and days 1 and 7($t(30) = -2.23$, $p<.05$), but not between days 0 and 1($p>.05$). For condition 3, where the objects were all semantically related, but were placed in random locations around the circle, there was a significant difference in error across test days ($F[2, 45] = 3.36$, $p<.05$), such that earlier test days had less errors (Day1: M=37.14[24.06]; Day2: M=49.62[34.41], Day3: M=71.31[50.13]). Planned post hoc analyses revealed a significant difference between days 0 and 7 ($t(30) = -2.46$, $p<.05$), but not between days 0 and 1 ($p>.05$), and days 1 and 7($p>.05$). The changes in error across days, especially between days 0 and 7 may suggest that forgetting is taking place over time. This may potentially result in a difference in contribution of the three error

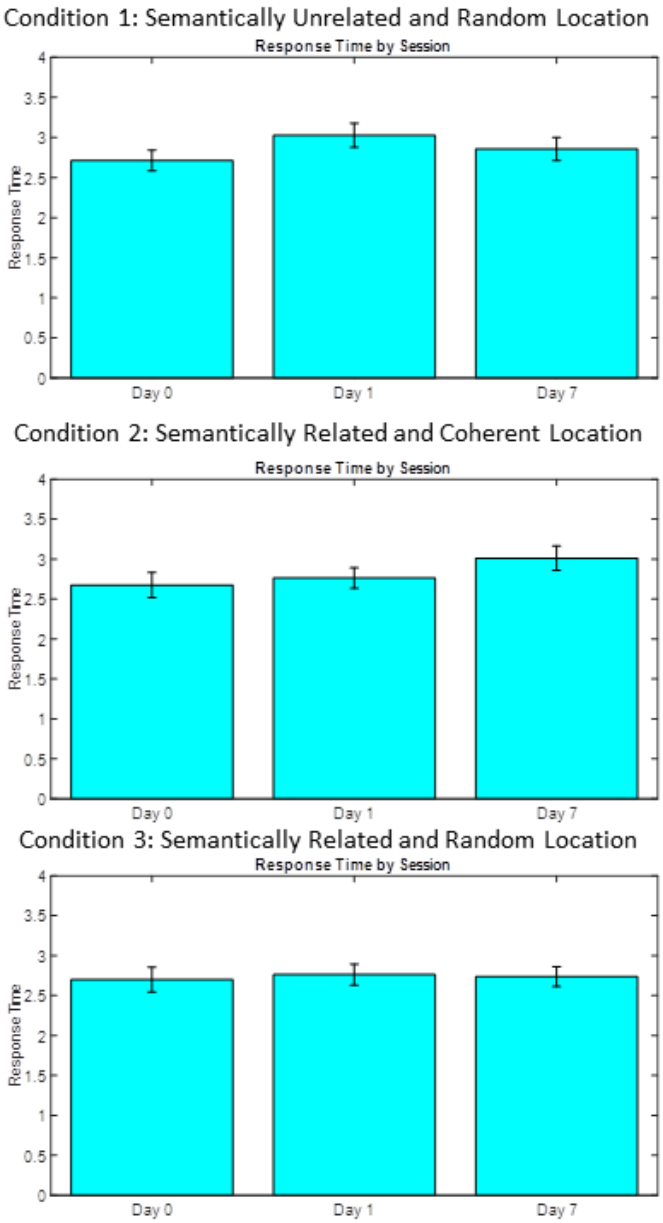

Figure 4.4 Response Times Graphs show mean response time by testing day for all conditions.

types (i.e. noisy correct responses, misassociations, and random guesses) over time. For example, there may be more random guessing during day 7 compared to day 0.

Response Time Analysis

Based on a difference in error performance between days 0 and 7 for each experimental condition, we tested to see whether response times varied across days. A potential explanation for a difference in response time, is the remember/know distinction. 'Remember' responses (i.e., recollected responses that are accompanied by details of the encoding episode) are thought to have a slower reaction time than 'know' responses (i.e., judgments of familiarity that are not accompanied by specific episodic information; Gimbel & Brewer, 2011). A difference in response time might indicate that participants varied in their remember/know responses over time. Participants may have made more 'remember' responses during earlier testing days, which have less error, and more 'know' responses during later testing days. Alternatively, if there are no differences in response time, then the source driving differences in error across days might be more nuanced than the remember/know distinction. A potential source might a difference in the contribution of the three error types (i.e. noisy correct responses, misassociations, and random guesses), as previous work has found no difference in response time as a function of error type (Lew et al, 2015). In this way, a misassociation, for example, does not take longer to make than a correct association. Neither does a correct response take longer to make than a random guess. Figure 4.4 present graphs of response times by testing day across conditions.

For all conditions, a series of analysis of variance tests revealed no significant difference in response time across days ($p>.05$). As a check for the quality of the stimulus (unbiased by object), we also assessed for differences in response times by object to make sure participants did not spend more time recalling some objects relative to others. Across all conditions, there was no difference in response times by objects ($p>.05$). This

lack of difference in response time suggests that differences in performance in this task were more nuanced than the remember/know explanation can detect, potentially lending support to the idea that the change in error across days is due to differences in the contribution of the types of error.

Contribution of Noisy Memories, Misassociations, and Random Guessing

To evaluate the contribution of the three error types to performance in each condition, we employed the Misassociations model (Lew et al., 2015). The Misassociations model, a variant of the standard remember-guess model (see chapter 3), is a finite mixture model that was fit to subject responses and used to estimate the probability of responses being misassociations and random guesses as well as the noise of memories. There were two stages to implementing the model. The first stage was to fit the mixture model to the data to learn the probabilities of making each error type and the second stage was to use maximum likelihood estimation to then evaluate how each component contributed to the RMSE observed in the task.

The mixture model assumes that are three types of responses that can be made in the recall task: a noisy *correct* response of the true target location, an *incorrect* response of a non-target location (i.e., a misassociation – all other locations studied in the task), or a random guess (a location in the task where no study object appeared). A noisy response of the target location is assumed to be distributed as a two-dimensional Gaussian (for x and y locations) centered on the true location of the target. A misassociated response to a non-target is also assumed to be distributed as a two-dimensional Gaussian, but is centered on the non-target locations. A random guess is assumed to be a sample drawn from a truncated Gaussian centered on the stimulus environment (in this case the circle)

and bounded by the edges of the circle. In this way, the mixture model provides estimates of three parameters: the probability of selecting a target location (p_T), the precision[9] (noise) of location memories (σ), and the probability of selecting a misassociated location (p_M). The probability of random guessing (p_R) is determined by the probabilities of selecting the target and making a misassociation, such that p_R = 1 - p_T - p_M. The mixture-model likelihood of reporting a location y is given by:

$$P(y|t) = p_T N(y|x_t, \sigma) + p_M \left(\frac{1}{n-1}\right) \Sigma N(y|x_i, \sigma) + (1 - p_T - pM) R(y), \qquad \text{Eq (2)}$$

where x_t is the true studied location for a particular trial *t*. $\Sigma N(y|x_i, \sigma)$ is the sum over all other studied locations (x_i). This allows p_M to be evenly split among the other studied locations. The total number of studied locations is denoted as *n*. $R(y)$ gives the likelihood of randomly guessing$y$, which is the mean location of a truncated normal bounded by the edges of the circle. When the standard deviation of the truncated normal is large, it closely resembles a uniform distribution. The mixture model was fit to each testing block using a Gibbs sampler. The sampler had a burn-in of 500 samples and after the burn-in it used 700 samples. There was no thinning procedure implemented. The initialized values chosen for the three parameters were: .33 for both the probability of selecting a target location and the probability of selecting a misassociated location, and 60 for the noise on location memories.[10]

[9] Although precision usually refers to inverse variance, in the Misassociations model, precision is used conceptually (not computationally) and therefore, refers to the standard deviation of location memories. There appears to be a lack of consistency in the use of precision across remember-guess models. The standard remember-guess model developed by Zhang & Luck, 2008 used the same terminology of precision to describe standard deviation. However, the mixture model of Bays, Gorgoraptis, Wee, Marshall, & Husain, 2011, use precision as the inverse standard deviation.

[10] The initialized parameter values, number of samples, and the sampler chosen were all consistent with what was used by Lew et al., 2015.

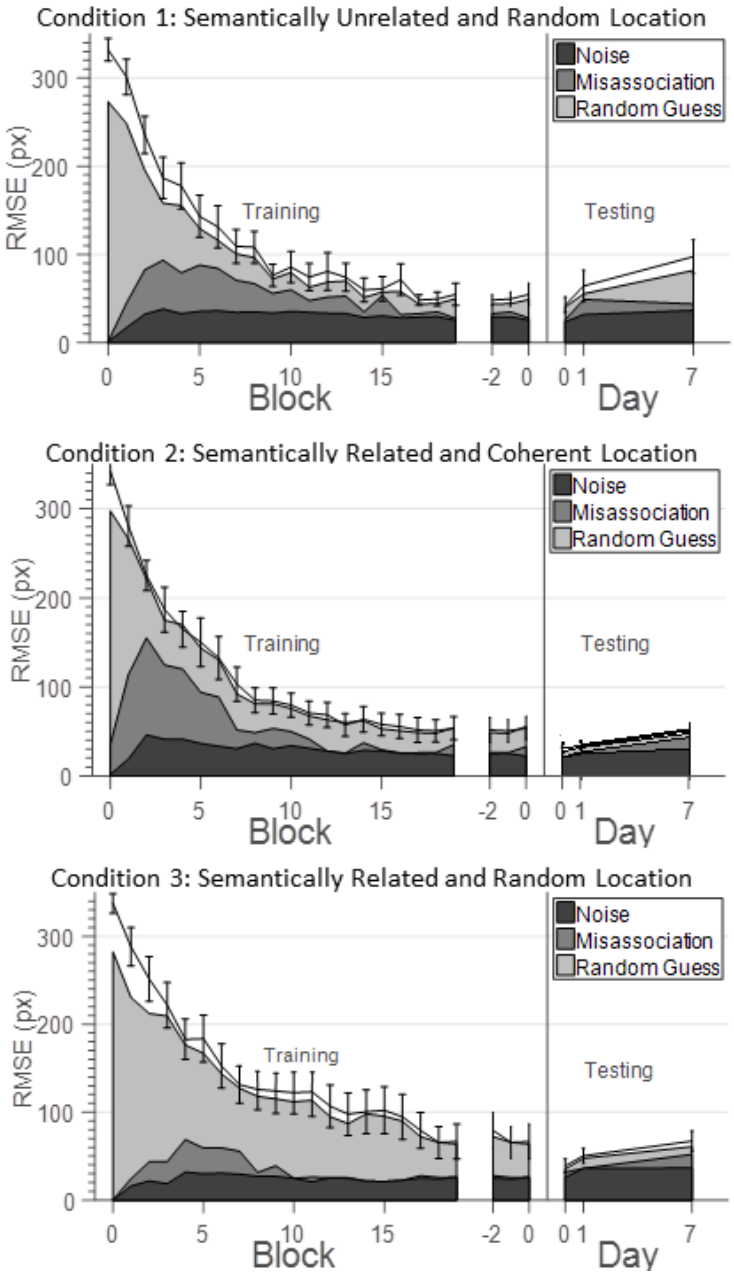

Figure 4.5. Error Partitions Graphs show learning and forgetting over time in each condition with errors partitioned based on their estimated source. The black line indicates subjects' raw root-mean-square error (RMSE; identical to Figure 4.3). Shading indicates the estimated errors due to noise from recalled locations, misassociations and random guessing.

During the second stage, maximum likelihood estimation was then employed to infer the contribution of each error type to RMSE, based on the estimate of noisy memory locations and the probabilities of making each error type learned from the model. Given

the parameter estimates, the models RMSE was calculated for each error type. In other words, for each response, the error due to the estimated standard deviation of location memories, the error due to the distance between the target location and the misassociated location (misassociation), and the error due to the distance between the target location and the center (random guess), was then calculated.

Most variants of the standard remember-guess model are employed to address questions regarding the fidelity of memory, resources of memory, and the capacity of memory, as detailed in Chapter 4. The Misassociation model, however, addresses the process of memory as it relates to competing information and the binding of information (e.g., object to location) that is implicit in all the other tasks modeled by remember-guess. In this way, this model presupposes a different contributor to long-term memory. Although this model is informative in deciphering the sources of error that contribute to performance across conditions, there are number of assumptions and limitations of this modeling framework as it pertains to characterizing data from this study. Given these limitations, drawing strict conclusions based on the model warrants caution. There are number of modifications and alternative models that could be implemented to evaluate the current model and/or potentially provide a better fit to the data. As a first pass, we employed the model strictly to parcel out error contributions, not to assess model fit or model assumptions. For a discussion of modeling assumptions and limitations, see Appendix section 2.

Figure 4.5 shows the model partitions of error for each condition. In condition 1 (semantically unrelated objects studied in random locations), it appeared that random guessing made the largest contribution to error at the onset of training, followed closely

by misassociations, and then noisy correct responses. As training progressed, there was a decrease in the amount of random guessing and misassociations, suggesting that participants were learning the object-location associations. During testing, the first day appeared to have the least amount of error and was comprised of mostly noisy correct responses, and a small portion of random guessing. The second day saw an increase in misassociations that tapered off over time. During the last day of testing, majority of the error was due to random guessing with a small contribution of noisy responses and even smaller contribution of random guessing. This trajectory of error contribution is consistent with the results of Lew et al. and suggests that participants learned the associations during training but forgot during testing as evidenced by the increase in random guessing.

Errors resulting from condition 2 training trials (semantically related objects studied in coherent locations) had a similar contribution of error types as condition 1. There was a large contribution of random guessing early on that decreased over training blocks. However, misassociations made a larger contribution to error in the earlier training blocks compared to condition 1. During testing, condition 2 deviated dramatically from condition 1 (and 3, see below). Critically important, there was relatively little contribution of random guessing in condition 2, even during the last day of testing for condition 2, compared to condition 1. This confirms our hypothesis that when associations are meaningful, the contribution that the sources of error makes to long-term memory changes. Specifically, using a strategy of random guessing disappears.

Training in condition 3 (semantically related object in random locations) was slightly different than conditions 1 and 2 in that majority of the error was comprised of random

guessing. There was a little contribution of noisy correct responses and even less contribution of misassociations. During testing, there was a greater contribution of random guessing than in condition 2, but less random guessing than condition 1. The results of the model partitions across conditions suggest that associations are better formed in long-term memory when they are semantically coherent and have basis in the observers' model of the real world. Arbitrary associations are less easy to form, and are likely to result in more random guessing. Taken together the difference in partitioning between conditions 1 and 2, these results confirm our hypothesis that memory behaves differently depending on the types of associations the system is tasked with recalling.

Decay Functions

To determine whether some aspects of memory were acquired more quickly than others, similar to Lew et al., we fit exponential decay functions to the parameter estimates (noisy location responses (here referred to as targets), misassociations, and random guesses). The decay function took the form:$B + (A - B)e_{\tau}^{-t}$, and quantified the speed with which these sources of error changed during training (learning). Parameters A and B of the exponential function indicated the initial and asymptotic values of the parameter estimates. Parameter τ reflected the time constant, where larger values indicate a slower rate of change and slower acquisition. Parameter *t* indicated the block number. Table 4.2 reports the mean exponential fits to each parameter for each condition.

The decay fits to the target parameter suggests that the initial proportion of targets (A of target) across conditions was relatively the same. However, the asymptote for the acquisition of targets (B of target) was much lower in condition 3, relative to conditions 1 and 2. Also, correct target associations were learned faster (τ of target) in condition 2

Table 4.2. *Mean Exponential Fits to Parameters*

	Target			Misassociation			Random guess			SD		
Condition	1	2	3	1	2	3	1	2	3	1	2	3
A	.04	.05	.04	.13	.22	.12	.93	.88	.95	42.92	32.63	50.78
B	.86	.90	.69	.09	.08	.04	.08	.03	.23	37.81	36.20	36.28
τ	2.57	2.18	1.89	1.64	1.52	1.72	2.16	1.71	2.18	.75	.73	.45

compared to condition 1, and fastest in condition 3. However, also note that the asymptote for condition 3 was much lower than condition 2, which might suggests that there were less targets acquired, and therefore, less time needed to learn those fewer targets.

Participants also made more misassociations initially in condition 2 compared to conditions 1 and 3, as evidenced by a higher *A* value in misassociations. In condition 2, participants made slightly less random guesses than in condition 1 and condition 3, as reflected in the lower *A* value for random guesses. Furthermore, participants had less initial error on location memories in condition 2 compared to condition 1 and 3 (lower *A* of SD). Across all conditions, the time constant (τ) for correct associations (target) was considerably larger than the imprecision of locations (τ of SD), which indicated slower learning of associations than accurate recall of exact locations.

The decay fits to the parameter estimates suggest that correct associations were learned fastest in condition 2, relative to the other conditions and there were more misassociations than random guesses. This pattern of training was consistent with the pattern of testing where most of the responses in this condition were classified as noisy correct responses and misassociations, but not random guesses.

Discussion

Long-term episodic memory encodes the associative relationships between pieces of information, such as the location of objects in space. In a real world scenario, this associative information equates to remembering the location of an airplane above an island. Previous research suggests that associations between information (i.e., objects to locations in space) can be difficult to form and retain over time (Lew et al., 2015). Recall of this information can result in three types of error: noisy correct responses, misassociations (pairing objects to the wrong locations), and random guesses. The contribution of these error types, however, might be mediated by the degree of meaningfulness between the associations. For example, objects that are associated to semantically coherent locations in space (e.g. airplanes in the sky), might be more quickly formed during learning and allow for more strategic or informed guessing when trying to retrieve associations from memory. People can use expectations for the prototypical associations of information to inform retrieval or guessing.

In this work, we explored the contribution of the three types of error in memory as a function of the meaningfulness of associations. We first assessed people's prior expectations for the associations of objects to locations in space and then used an established paradigm for assessing misassociations to measure memory for associations (Lew, et al., 2015). In the task assessing expectations (Experiment 1), we simply asked people to place objects where they normally expect to find them. This procedure revealed a strong subjective agreement in the placement of objects that were semantically related to the center object. Objects that prototypically belonged above the center object were consistently placed in that area and objects the prototypically belonged below the center

object were placed below. Objects that did not have a prototypical location assignment relative to the center object (i.e., random objects) did not reflect the same level of consistency in placement. Overall, the pattern of behavior observed in the prior knowledge task revealed that people have clear expectations for the placement of objects to locations. These expectations, in turn might impact how associations in an experimental task are formed and stored in long-term memory.

In the learning and memory tasks (Experiment 2 and 3), observers were trained on the location of objects in a circle, and then completed three cued recall tasks over the course of 7 days. Importantly, we manipulated the degree of meaningfulness of the object-location associations. The results illustrated that when the associations were the least meaningful (condition 1), there was an increasingly large contribution of random guessing over time. When the associative information was the most meaningful (condition 2), there was only minimal random guessing, even after a 7 day retention interval. When the associations were moderately meaningful (condition 3), there was considerably more forgetting than the most meaningful associations, but less forgetting than the least meaningful associations. Taken together, the results from these experiments demonstrated that the standard finding of high rates of random guessing (i.e. guessing relatively uniformly across the stimulus space) hinges on the stimuli with which the memory system is tasked with storing and retrieving.

The difference in finding between conditions 1 and 2 regarding the formation of associations and the contribution of random guessing to recall can be reconciled by understanding the nature of the to-be-remembered associative information. Arbitrary associations might be difficult to form and store over time. This can lead to individuals

resorting to guessing randomly across the stimulus environment. People have clear expectations for object locations; however, those expectations might not be useful in a task where the target-cue pairing is intentionally random. In this case an individual could only recall the information, use other task relevant information (misassociations), or guess. On the other hand, when associations are meaningful, people can use their prior expectations to fill in for noisy memories. The influence of prior knowledge can be observed as informed guessing. Instead of guessing uniformly across the space, people might guess using the hemisphere they expect to find the object located. The Misassociations model, in its current form, does not parameterize this type of guessing. However, the model could be modified to have an added mixture component to capture this type of informed guessing. Alternatively, based on the output of the model in the fully associated condition (no random guessing), it is possible that people did not use a guessing strategy at all, (if one considers making misassociations a form of guessing). Recall errors from the fully associated condition were classified as noisy correct responses and misassociations. To confirm the possibility of no random guessing in the fully associated condition, a model could be implemented that removes the guessing component altogether. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values can be calculated for this alternate model and compared to values of the full model.

Importantly, when the stimulus structure closely resembles that of the natural environment and memory is noisy, individuals either missassociate information or use prior knowledge as a strategy to inform retrieval, but they do not randomly guess. This

finding has significant implications for how working and long-term memory are understood and transformative potential for current models of memory.

Appendix

Section 1 Location Responses

The results from the memory tasks across conditions are presented as root mean square error (RMSE) in the main text. However, this error measure obscures some of interesting trends in the data, such as whether the locations that participants recalled reflected the differences in study locations across conditions. Here, we present figures from the recall results in each condition. Figure 4.6, shows the study locations for three of the objects in the fully associated conditions across participants (left panel). It also shows both the studied (black square) and recalled locations (red circle) for those objects across the three testing days. The first row contains the study and recall locations from an object that was always presented above the island (e.g. cloud). The second row shows the locations for an object presented in line with the island (e.g. jet ski) and the third row shows an object presented below an island (e.g. shark). The graphs suggest that participants were sensitive to the study location of the object and there were only a small number of responses that were given that fell well below the true studied locations (in the southern hemisphere). A similar pattern was observed for the objects that fell online with island as well as the objects that fell below the island. This pattern was observed even after the 7-day retention interval (testing day 3). Although only one object from each category is presented here, the results were consistent across all objects within these categories.

In the partially associated condition (Figure 4.7) the locations of these same objects were spread throughout the circle. This finding may reflect people's sensitivity to the studied locations, relative to their prior expectations for where these objects belong. If people where strictly using their prior expectations, especially during the last testing session, there might be more responses reflecting the semantically correct hemispheres. However, this pattern of responding was not observed. Similar to the partially associated condition, in the fully random condition (Figure 4.8), where the objects were not associated pre-experimentally to a particular location, the response locations were dispersed throughout the circle.

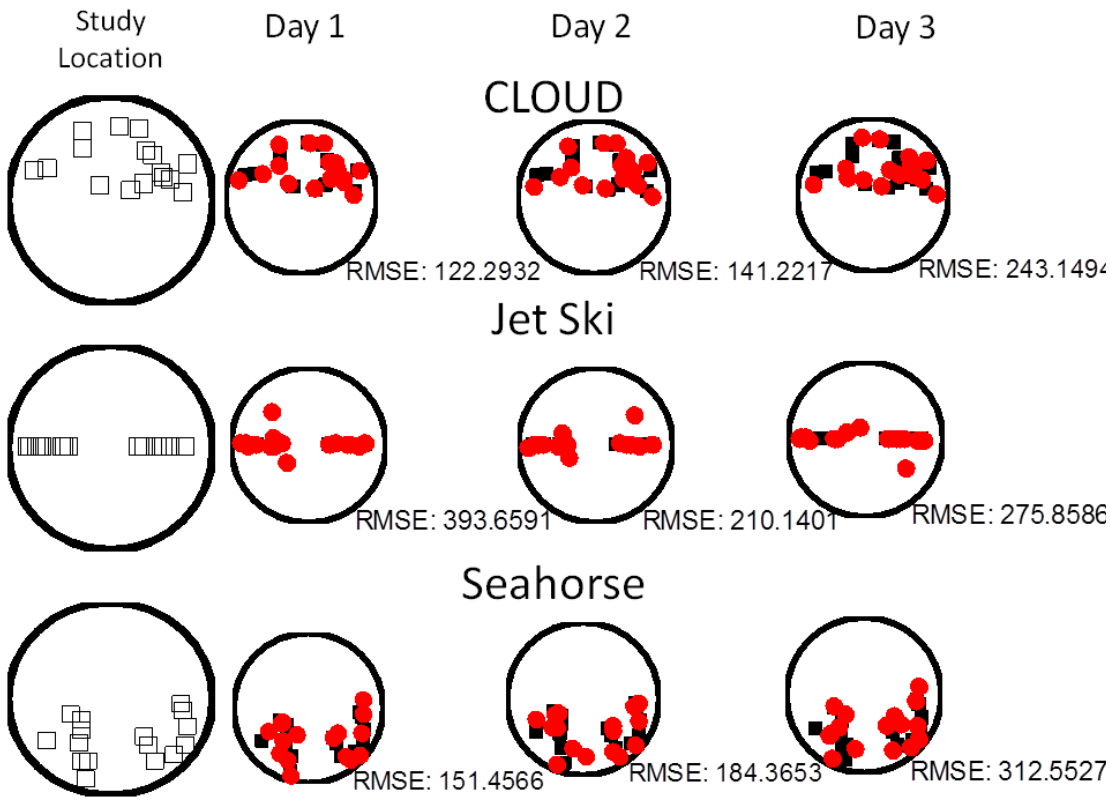

Figure 4.6. Graphs of study and response locations across days for objects from the fully associated condition

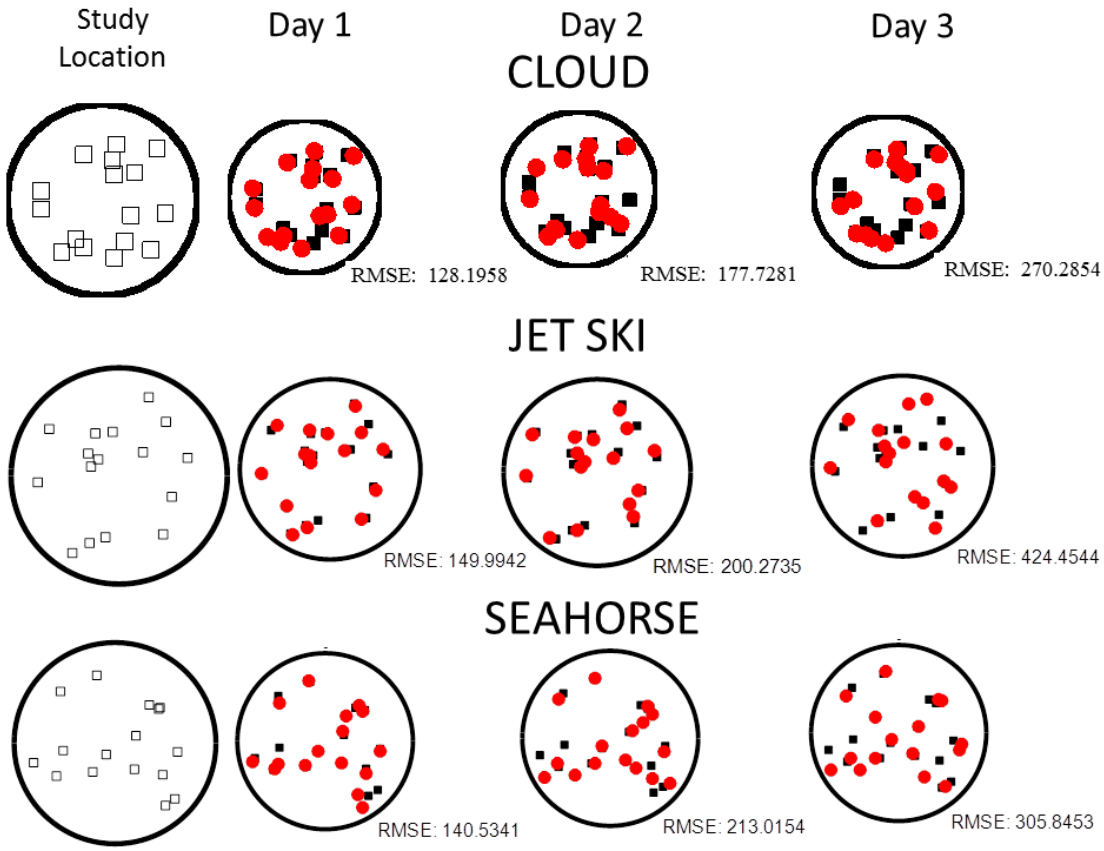

Figure 4.7. Graphs of study and recall locations across participants in the partially random condition.

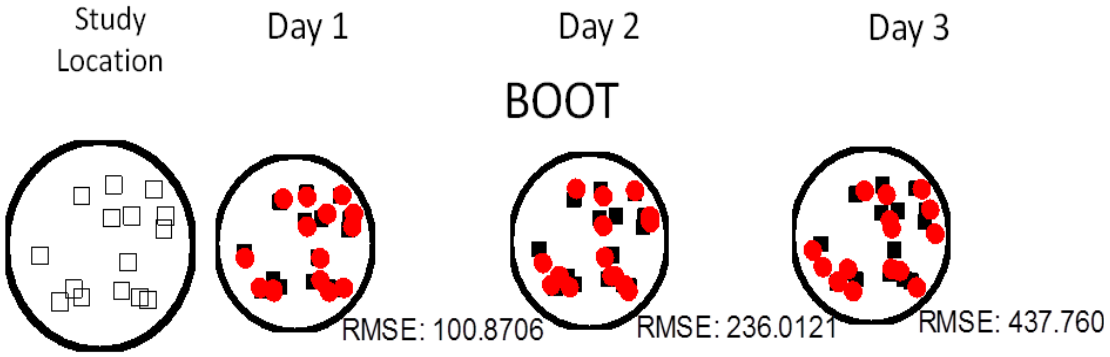

Figure 4.8. Graphs of study and recall locations across participants in the partially random condition.

Section 2 Misassociations Model

There are number of assumptions and limitations of the Misassociation model. First, the model assumes that the noise on the target location is equal to the noise on all non-target locations, and therefore only estimates a single precision parameter for location memories. This assumption is also a limitation of the model because it is possible that there is variability in precision for locations associated with incorrect objects relative to correct objects. However, the authors of the model argue that memory for locations and associations are stored separately, so the precision of the location is theoretically independent of the associations. In this way, the precision of the location memory is the same regardless of the association being correct. This is the justification for assuming a single parameter for noise on location memories. To address this issue, an alternative model that adds another parameter to allow for a difference in precision for correct and misassociated objects could be implemented. Model comparisons between this alternative model and the single noise parameter model can be performed (as long as the test acknowledges the difference in the number of parameters). If the single noise model provides the superior fit to the data, then the assumption might be justified.

Another assumption of the model is that random guesses are drawn from a truncated normal distribution centered in the stimulus environment (i.e. the center of the circle). This assumes that when individuals randomly guess, they use the center of the circle as opposed to guessing uniformly across the circle. This assumption deviates from other instantiations of the remember-guess paradigm where random guessing is parameterized as a draw from a uniform distribution over all response options (Bays, Catalao, & Husain, 2009; Bays, Gorgoraptis, Wee, Marshall, & Husain, 2011; Brady et al., 2013; Zhang &

Luck, 2008). This difference poses a theoretical concern in the interpretation of what a random guess is. In the Misassociations model, the random guess is not a 'random guess' (in the traditional characterization sense) because it uses some information about the task, namely the location of the center of the circle to inform the response. It could be argued that the use of this type of information is not a random guess, but an informed guess. However, when variability in the truncated Gaussian is relatively high, the distribution approaches a uniform. This might be the rationale for the authors of the model choosing to parameterize random guessing in this way. Unfortunately, this further complicates the issue of what working and long-term memory researchers mean when they say guessing randomly.

A third assumption, and arguably the most alarming given the experimental manipulation (type of associations), is that the model assumes independence in the probabilities of error type. By assuming independence, the model assigns equal probabilities to the error types across trials. However, it is conceivable that when participants are performing in the task, they may be aware that they correctly placed an object on an earlier trial, which can then change the probabilities of making the other types of error, because participants now have to consider fewer locations. This issue is even more salient in the fully associated condition, where participants might have learned the number of objects that appeared above or below the island, adding another layer of location dependence.

In its current form, the model does not use built-in assumptions about possible changes in probabilities over trials, and therefore does not account for interactions among targets. In other studies using the remember-guess paradigm, the targets are usually

sampled independently of one another,(e.g. colored squares in a circle- Bays, Catalao, & Husain, 2011; orientation of colored lines – Bays, Gorgoraptis, Wee, Marshall, & Husian, 2011), and an interaction of targets is never addressed in the model. For example, Bays and colleagues (2011) used a finite mixture model similar to the one used in this paper (classifies error into the three types). However, in the task, participants studied the orientation of colored lines and importantly, the orientations were sampled independent of one another. While the orientations are selected at random and independent, in the task described in this paper, there is a constraint that the studied locations of objects cannot overlap with one another which indicates some dependence of target locations and potential non-independence at retrieval. A model that is more suitable for analyzing performance in this task would need to take into account the potential interaction between targets as decisions are made across trials.

A model that is sensitive to the lack of independence might be revised in a number of ways. For instance, the Misassociation term in the model is a sum over all locations that are not the target. If on the first trial, a participant correctly places an object to a location, this location can then be ruled out as a potential misassociation location for later trials, effectively changing the probability of misassociations. The only issue with this approach is that it is hard to know a priori if this is how participants truly behave. It is possible that even when participants believe they have placed an object correctly, when they see another object that is potentially similar (e.g. it also belongs above an island), they can still missassociate to the location that once was deemed correct. Alternatively, a model would need a substantial number of additional parameters to capture the relationship and changes of error types across trials. Furthermore, behavior might be different across

individuals, thus a model of individual differences might also be warranted. Given the issue of assumptions with the model and the experimental manipulations in the task, alternative models need to be considered.

The goal of implementing the model was to evaluate if the contribution of the three error types change across experimental conditions, and was primarily intended as an aid to the empirical study. Therefore, it would also be informative to evaluate how this particular model compares to alternative models in terms of explaining the data. For example, it was speculated that random guessing would be absent from the fully associated condition. A model with the random guessing component removed, but still contained misassociations and noisy correct responses (a two-term model), could be implemented and compared to the full model. If the two-term model provides a better fit to the fully associated condition, this might provide stronger evidence the nature of the associations in the fully associated condition resulted in less random guesses.

Furthermore, this study discussed in this chapter attributes the lack of random guessing (guessing uniform over the stimulus space) in the fully associated to the fact that people have prior expectations for the locations of objects and used that information in the task. However, the current model does not parameterize the role of this contribution. To address this, a variant model could be implemented that adds or changes what is considered guessing the task. If people are using their expectations, they might not guess uniformly around the circle. Instead, they might sample from a truncated distribution that is not only bounded by the edges of the circle but also by the edges of the hemisphere for which the studied object belong (i.e. either above or below the island). This model can either add this type of random guessing or change the current random guessing

component to reflect this behavior. In this way, using prior expectations is assumed to be informed guessing as opposed to a feature of memory. Again, model comparisons for the two types of guessing components could be implemented to evaluate this contribution.

Chapter 6: Conclusion

A fundamental question regarding long-term memory (LTM) is what happens to information over time. There are several theoretical frameworks spanning working and LTM that have been employed to address this question. A particularly popular framework is the remember-guess paradigm which characterizes errors in WM as either noisy memory traces or random guesses. In this paradigm, it is assumed that once memory traces become too noisy, or are no longer retrievable, people resort to random guessing (Brady, Konkle, Gill, Oliva, & Alvarez, 2013). While it has been demonstrated that LTM has an impressive storage capacity (Brady, Konkle, Alvarez, & Oliva, 2008), work within this paradigm suggests that the system struggles to form and store associative information between memories, resulting in fragile associations (Lew, Pashler, & Vul, 2015). In this way, information is either noisily recalled, misassociated to other studied content, or no longer retrievable.

What often goes unmentioned is the role of prior knowledge in memory. People have a rich database of knowledge that they bring to the task of reconstructing events from memory (Hemmer & Persaud, 2014). In Chapter 2 of this dissertation, we learn that people have strong subjective agreement on expectations for information that is consistent with the environment. We also learn that memory is biased and can reflect those expectations, especially when episodic information is noisy. In Chapter 3 we learn that the biases in memory also persist across cultures, suggesting that the use of prior knowledge might be a general mechanism for reconstructing events from memory. This relationship between prior expectations and episodic memory is well fit by a generative rational model under the simple assumption that people combine expectations for the

statistical regularities in the environment with noisy episodic representations to produce recall.

The benefit of prior knowledge is most notable when the stimulus environment is ecologically valid and consistent with the statistical regularities of the real world. Prior knowledge is used to fill in episodic information when memory traces are noisy or incomplete. Furthermore, the influence of prior knowledge has also been demonstrated across a number of other cognitive domains, including attention (Kim & Rehder, 2011), functional memory capacity (Ricks & Wiley, 2009), and perceptual categorization (Huttenlocher, Hedges, & Vevea 2000). In this way, prior knowledge can facilitate storage and retrieval of episodic information over time and the role of prior knowledge should not be ignored in theories and models of memory.

In Chapter 4 we draw on analytical practices from the remember-guess framework and Bayesian framework (as described in Chapter 2), to extend the memory study from Chapter 2 and perform model comparisons. The extension of the free recall task also revealed a bias in memory toward categorical expectations. Importantly, this bias was observed regardless of whether or not participants were cued to think about the category (i.e., asked to explicitly label the studied colors).

The aggregate data from this task appeared to have a high rate of random guessing. However, when partitioned by lag (i.e., the number of intervening trials between study and test), immediate memory mirrored perception in its high fidelity, but with increasing lag, intermediate memory appeared to be more complex, and at longer lags recall appeared to be a mixture of episodic information, and guessing. Performance at intermediate lags, consistent with the Bayesian assumption, might reflect the influence of

category knowledge on noisy episodic representations. Three models were then implemented, including the standard 'Remember-Guess' model, a variant of this model, and a Bayesian memory model. Relative to the data, the variant of the standard model which posits two states of fidelity (high fidelity and low fidelity) for memory provided a superior fit. Although this model fits the data, it was agnostic about what constitutes the low fidelity memory state and therefore did not capture the regression to the mean pattern born out in the data. In fact, the Bayesian model was the only model to capture this regression behavior. The results from this study suggest that memory is not simply a combination of remembering and guessing, but other factors may influence performance.

Chapter 5 presents research from a new study that evaluated the potential contribution of another factor to long-term memory, namely misassociations. Previous work has suggested that observers recall target information with noise, misattribute target and cue information, or guess randomly (Lew, Pashler, & Vul, 2015). The research presented in Chapter 5 extended the findings from this earlier study by manipulating the stimulus environment to further evaluate the role prior knowledge and misassociations in memory. The experiments in this study assessed 1) people's prior knowledge and expectations for associative information, 2) cued recall for random objects in random locations, 3) associated objects in meaningful locations, and 4) associated objects in random locations. The results revealed that memory for meaningful associations, relative to random associations, did not result in a high contribution of random guessing. Lew et al.'s Misassociations model (Lew et al., 2015) which classifies the errors in memory as noisy memory, misassociations (misbinding of target and cue), and random guesses was fit to the data.

The Misassociations model fit to the random associations data (i.e., unrelated objects in random locations) demonstrated that errors in memory were comprised of a large portion of random guessing and misassociations, while the model fit to the meaningful associations (i.e., associated objects in meaningful locations) demonstrated that a large portion of errors in memory resulted from noisy memories, misassociations, but *not* random guessing. The lack of random guessing could either indicate that when recalling meaningful associations, guessing is not a strategy that is employed or that instead of guessing randomly, people use prior knowledge to guess with relevant information. The latter explanation might explain the presence of misassociations in the fully associated condition. The results of this study support the hypothesis that there is little to no random guessing in LTM for semantically associated, ecologically valid stimuli and prior knowledge may underlie this finding.

The overarching message of the work discussed in this dissertation is that people have a wealth of knowledge that they bring to the task of remembering and this knowledge often influences the reconstruction of events from memory. The influence of prior knowledge and expectations is further amplified when the stimulus environment in the memory task reflects information from the real world. Theories of long-term memory, especially those that are extensions from working memory should acknowledge the role that prior knowledge plays when designing experiments and drawing conclusions about memory as a reconstructive process.

References

Baddeley, A.D. & Scott, D. (1971). Short-term forgetting in the absence of proactive interference *Quarterly Journal of Experimental Psychology, 23,* 275–283.

Bae, G.Y., Olkonnen, M., Allred, S., & Flombaum, J. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General, 144,* 744–763.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133,* 83–100.

Bartlett, F.C. (1932). Remembering: A study in experimental and social psychology, Cambridge, England: Cambridge University Press.

Bays, P. M., Catalao, R.F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9,* 7–11.

Bays, P.M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision, 11,* 6-21.

Bays, P.M., Wu, E.Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia, 49,* 1622–1631.

Belli, R.F. (1988). Color blend retrievals: Compromise memories or deliberate compromise responses. *Memory & Cognition*, 16, 314-326.

Berlin, B. & P. Kay (1969). Basic Color Terms: Their Universality and Evolution. *Berkeley: University of California Press*.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*, 14325–14329.

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science, 24*, 981–990.

Devkar, D. & Wright, A. (2016). "Event-based proactive interference in rhesus monkeys". *Psychonomic Bulletin & Review*, 1-9.

Donkin, C., Nosofksy, R., Gold, J., & Shiffrin, R. (2014). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychological Bulletin and Review, 21*, 2–11.

Eckstein, M. P., Abbey, C. K., Pham, B. T., & Shimozaki, S. S. (2004). Perceptual learning through optimization of attentional weighting: Human versus optimal Bayesian learner. *Journal of Vision*, 4(12), 1006-1019.

Epstein, R.A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in Cognitive Sciences*, *12(10),* 388-396.

Galleguillos, C. & Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding, 114(6),* 712-722.

Goldstone, R. (1995). Effects of Categorization on Color Perception. *Psychological Science*, 6(5), 298-304.

Fawcett, J. & Lawrence, M., & Taylor, T. (2016). The representational consequences of intentional forgetting: Impairments to both the probability and fidelity of long-term memory. *Journal of Experimental Psychology: General, 145*, 56-81.

Fischer, J. & Whitney, D. (2014). “Serial dependence in visual perception”. *Nature Neuroscience, 17,* 738-743.

Fraley, C. & Raftery, A.E. (2006). “MCLUST version 3 for R: normal mixture modeling and model-based clustering”. (Technical report no. 504, Department of Statistics, University of Washington).

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). Bayesian data analysis. *Boca Raton, FL: Chapman & Hall.*

Gimbel, S. & Brewer, J. (2010). Reaction time, memory strength, and fMRI activity during memory retrieval: Hippocampus and default network are differentially responsive during recollection and familiarity judgments. Cogn*itive Neuroscience, 2,* 19-26.

Griffiths, T.L. & Tenenbaum, J.B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*, 767–773.

Hemmer, P. & Persaud, K. (2014). Interactions between categorical knowledge and episodic memory across domains. *Frontiers in Psychology, 5*, 1-18.

Hemmer, P., Persaud, K., Kidd, C., & Piantidosi, S. (2015). Shifting priors: Evaluating the crosscultural influence of color expectations on episodic memory. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Hemmer, P. & Steyvers, M. (2009a). Integrating Episodic Memories and Prior Knowledge at Multiple Levels of Abstraction. *Psychonomic Bulletin & Review, 16*, 80–87.

Hemmer, P. & Steyvers, M. (2009b). A Bayesian Account of Reconstructive Memory. *Topics in Cognitive Science, 1,* 189–202.

Hemmer, P., Tauber, S., & Steyvers, M. (2015). Moving beyond qualitative evaluations of Bayesian models of cognition. *Psychonomic Bulletin & Review, 22,* 614–628.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in establishing spatial location. *Psychological Review, 98,* 352–376.

Huttenlocher, J., Hedges, L.V., & Vevea, J.L. (2000). Why Do Categories Affect Stimulus Judgment? *Journal of Experimental Psychology, 129,* 220–241.

Jern, A. & Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology.* 66(1), 85-125.

Kim, S & Rehder, B (2011). How prior knowledge affects selective attention during category learning: An eyetracking study. *Memory & Cognition, 39,* 649-665.

Lew, T.F., Pashler, H.E., & Vul, E. (2015). Fragile associations coexist with robust memories for precise details in long-term memory. *Journal of Experiment Psychology: Learning, Memory, and Cognition. Advance Online Publication.*

Lewandowsky, S., Oberauer, K., & Brown, G. (2009). No temporal decay in visual short-term memory. *Trends in Cognitive Sciences, 13,* 120–126.

Loftus, E. F. (1977). Shifting human color memory. *Memory & Cognition,* 5, 696-699.

Meo, T. & McMahan, B. & Stone, M. (2014). Generating and Resolving Vague Color References. In Verena Rieser and Philippe Muller (Eds.) *Proceedings of the 18th Workshop in Semantics and Pragmatics of Dialogue* (pp 107-115). Edinburgh, Scotland

Mitterer, H. & de Ruiter, J. P. (2008). Recalibrating color categories using world knowledge. *Psychological Science,* (16), 629-634.

Neath, I. & Brown, G. (2012). Arguments against memory trace decay: a SIMPLE account of Baddeley and Scott. *Frontiers in Psychology, 3,* 1–3.

Palmer, S. & Schloss, K. (2010). An ecological valence theory of human color preference. *Proc. Natl. Acad. Sci. USA, 107,* p.8877-8882.

Paramei, G.V. (2005). Singing the Russian blues: An argument for culturally basic color terms. *Cross-Cultural Res*. 39, 10–38.

Persaud, K. & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1162–1167). Quebec City, CA: Cognitive Science Society.

Portrat, S., Barrouillet, P., & Camos, V. (2008). Time-related decay or interference – based forgetting in working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1561–1564.

Ricks TR. & Wiley, J (2009). The influence of domain knowledge on the functional capacity of working memory. *Journal of Memory and Language, 61,* 519-537.

Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2004). The development of color categories in two languages: A longitudinal study. *Journal of Experimental Psychology: General, 4,* 554–571.

Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology, 50,* 378–411.

Roberson, D., Davies, I., & Davidoff, J. (2000). Colour categories are not universal: Replications and new evidence from a Stone-Age culture. *Journal of Experimental Psychology: General, 129,* 369–398.

Shiffrin, R.M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 64*, 583-616.

Steyvers, M., & Griffiths, T. L. (2008). Rational Analysis as a Link between Human Memory and Information Retrieval. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects from Rational Models of Cognition* (pp. 327–347). Oxford: Oxford University Press.

Steyvers, M., Griffiths, T.L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10*, 327–334.

Stickles, E. and Regier, T. (2014). The relation of color naming and the environment. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Suchow, J. W., Brady, T.F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision, 13*, 1–8.

Todorovic, D. (2010). Context effects in visual perception and their explanations. *Review of Psychology, 17*, 17-32.

van den Berg, R., Shin, H., Chou, W., George, R., & Ma, W.J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academies of Science, 109,* 8780 –8795.

van der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica, 1,* 45-56.

Winawer, J., Witthoft, N., Frank, M.C., Wu, L., and Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academies of Science, 104,* 7780–7785.

Zhang, W., & Luck, S.J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453,* 233–235.