UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Flexibility and Biases in Cognitive Control and Categorization

Permalink
https://escholarship.org/uc/item/73p908dh

Author
Xu, Jing

Publication Date
2011

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Flexibility and Biases in Cognitive Control and Categorization

By

Jing Xu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Thomas J. Griffiths, Co-Chair
Professor Richard B. Ivry, Co-Chair
Associate Professor Terry Regier

Fall 2011

Flexibility and Biases in Cognitive Control and Categorization

Copyright 2011
by
Jing Xu

Abstract

Flexibility and Biases in Cognitive Control and Categorization

by

Jing Xu

Doctor of Philosophy in Psychology

University of California, Berkeley

Associate Professor Thomas L. Griffiths, Co-Chair

Professor Richard B. Ivry, Co-Chair

How do we maintain balance in our daily life, yet at the same time, adapt to changes in the environment? In the research presented here, I attempt to address this question in two ways. First, I examine the flexibility and biases associated with processes of inhibitory control. Second, I explore biases in categorization. Both lines of research are grounded in a probabilistic view of the world.

In the first line of research, I examine two aspects of inhibitory control. In one study, I ask if people can selectively inhibit their on-going actions. Using transcranial magnetic stimulation (TMS) to probe corticomotor excitability, I examine if inhibition can be selectively directed to a specific muscle, or if these inhibitory commands also modulate other muscles. In a second experiment of this study, I trained participants to selectively stop. I develop a Sampling-Bias model to analyze these data and find that costs associated with selective stopping are likely due to a statistical sampling bias. In the second study on inhibitory control, I examine the effects of intrinsic fluctuations in motor excitability on the dynamics of inhibitory control signals. Using TMS applied prior to the start of a trial as a measure of excitability, I observed an interaction of intrinsic and extrinsic effects, with the evidence suggesting that the dynamics of the control signal outweigh the effects of intrinsic fluctuations.

In the second line of research, I use a novel experimental paradigm and computational model, iterated-learning, to explore how biases influence categorization. In one study, I test the prediction that iterated-learning converges to commonly shared biases, applying this to Bartlett's classical serial-reproduction paradigm. In a second experiment, I use this paradigm to simulate the cultural evolution of color categorization, seeking to understand if an iterated-learning process, coupled with common perceptual and learning biases, will converge to the linguistic universals found in the World Color Survey.

Both lines of research underscore a common theme: Understanding human cognition requires consideration of our mental biases.

To my parents, for their constant support and unconditional love

Contents

List of Figures **v**

List of Tables **x**

1 Introduction **1**
1.1 Flexibility and biases in cognitive control: Hardware or software? 1
1.2 Surfing the waves of fluctuations in cognitive control: How do unknown biases influence us? 3
1.3 Positioning ourselves in a probabilistic world: Empirical and computational approaches 4
1.4 Representing the world: what are the biases in our mind and in the world out there? 5

2 Flexibility and Specificity in Inhibitory Control **7**
2.1 Introduction 7
2.2 Materials and methods 8
2.2.1 Participants 8
2.2.2 Experiment 1 8
2.2.3 Experiment 2 12
2.3 Results 16
2.3.1 Experiment 1 16
2.3.2 Experiment 2 20
2.4 Discussion 23
2.4.1 Psychological issues on selective stopping 23
2.4.2 When is global and when is local? 26
2.4.3 Neural mechanisms of the global and local inhibition 27
2.4.4 Conclusion 27

3 Intrinsic Fluctuation in Inhibitory Control **29**
3.1 Introduction 29
3.2 Materials and methods 30
3.2.1 Participants 30

3.2.2 Experimental task . . . 31
3.2.3 Assessment of cortical excitability . . . 32
3.3 Results . . . 34
3.3.1 Experiment 1 . . . 34
3.3.2 Experiment 2 . . . 37
3.3.3 Experiment 3 . . . 39
3.4 Discussion . . . 41
3.4.1 Interplay of the extrinsic and intrinsic activities: the case of failed stop . . . 42
3.4.2 Sequential effects in the stop-signal task . . . 42
3.4.3 Neural substrates of intrinsic fluctuation . . . 43
3.4.4 Conclusion . . . 43

4 A Rational Analysis of Serial Reproduction 44
4.1 Introduction . . . 44
4.2 A Bayesian view of serial reproduction . . . 45
4.2.1 Reconstruction from memory . . . 45
4.2.2 Serial reproduction . . . 46
4.3 Serial reproduction of one-dimensional stimuli as autoregression . . . 48
4.3.1 Testing the model predictions . . . 50
4.4 Experiment 1: Between-Subjects serial reproduction . . . 51
4.4.1 Method . . . 51
4.4.2 Results and discussion . . . 52
4.5 Experiment 2: Within-Subjects serial reproduction . . . 54
4.5.1 Method . . . 55
4.5.2 Results and discussion . . . 55
4.6 Experiment 3: Serial reproduction with no training on priors . . . 58
4.6.1 Method . . . 59
4.6.2 Results and discussion . . . 59
4.7 Experiment 4: Serial reproduction with bimodal distributions . . . 59
4.7.1 Method . . . 61
4.7.2 Results and discussion . . . 61
4.8 General discussion . . . 63
4.8.1 Connections to other research . . . 64
4.8.2 Limitations and future directions . . . 66
4.8.3 Conclusion . . . 67

5 Reproducing Color Term Universals 68
5.1 Introduction . . . 68
5.2 Background on iterated learning . . . 71
5.3 Materials and methods . . . 72
5.3.1 Participants . . . 72

5.3.2 Stimuli 72
5.3.3 Procedure 72
5.3.4 Comparing partitions 73
5.4 Main results 74
5.5 Additional analyses 76
5.5.1 More detailed analysis of random chains. 76
5.5.2 Assessing convergence. 77
5.5.3 Fixing the number of terms. 77
5.5.4 Other methods for comparing partitions. 79
5.6 Conclusion 82

6 Conclusion: From the self to the world **84**

References **85**

List of Figures

2.1 Stop-signal task procedure and TMS timeline used in Experiment 1a (A), and Experiment 1b (B). C. Stop-signal task procedure used in Experiment 2. 10
2.2 Horse-race model and the sampling-bias hypothesis for the residual cost. A. Graphic illustration of the horse-race model for stop-signal paradigm. B. Graphic illustration of the sampling-bias hypothesis. C. Graphic illustration of the re-programming hypothesis for stopping cost. 14
2.3 Sampling Bias Model's simulation to illustrate the case when sampling-bias cost can fully account for the behavioral cost (A-D), and the case when a restart-cost exist in the data so that the sampling-bias cost can not fully account for the behavioral cost (E-H). A, E. Simulated Go, Failed Stop (FS), and Successful Stop (SS) Foot RT data using normal distributions. B, F. stopping function $P(stop|footRT)$ using logistic regression to fit the FS and SS data sets shown in (A & E). C, G. Normal fits of a new set of samples drawn from the Go Distribution in (A & E), and FS and SS distributions classified using the stopping function in (C & F). D, H. Cost in the simulated data (A & E) compared to the modeled Sampling-Bias cost. 15
2.4 Percentage change in the amplitude of MEPs in four functional classes of muscles in Experiment 1a. A. On a go trial compared with baseline; B. On a successful-stop trial compared with baseline; C. On a successful-stop trial compared with go. . . . 18
2.5 Proportion change in the amplitude of MEPs in left FDI, ADM, and FDI and ADM combined in Experiment 1b, on Go trials, Successful Stop trials relative to the two baselines combined, and on Successful Stop trials relative to Go trials. 20
2.6 Non-stopping effector, foot RT and selective-stopping cost data for experiment 2. A. Foot RTs by block. B. Cost by day. C. Estimated residual cost with Sampling-bias Model 1, compared to behavior cost. 22

3.1 Stop-signal task procedure and TMS timeline used in A. Experiment 1 and 3a, within-hand task. B. Experiment 2 and 3b, between-hand task. 32
3.2 Proportion change in the amplitude of MEPs in four functional classes of muscles in Experiment 1. A-D. Proportion of MEP change preceding Failed Stop (FS) and Successful Stop (SS) compared to Go trials. E-F. Proportion of MEP change after Failed Stop (FS) and Successful Stop (SS) compared to Go trials. 36

3.3 Proportion change in the amplitude of MEPs in two functional classes of muscles in Experiment 2. A-B. Proportion of MEP change preceding Failed Stop (FS) and Successful Stop (SS)compared to Go trials. C-D. Proportion of MEP change after Failed Stop (FS) and Successful Stop (SS) compared to Go trials. 38
3.4 Proportion change in the amplitude of MEPs in four functional classes of muscles in Experiment 3. A-D. Data for within-hand task: proportion of MEP change preceding Failed Stop (FS) and Successful Stop (SS) compared to Go trials (A & B), and proportion of MEP change after Failed Stop (FS) and Successful Stop (SS) compared to Go trials (C & D). E-F. Data for between-hand task: proportion of MEP change preceding Failed Stop (FS) and Successful Stop (SS) compared to Go trials (E), and proportion of MEP change after Failed Stop (FS) and Successful Stop (SS) compared to Go trials (F). 40

4.1 Model simulations for three types of behavior. Panels (a)-(c) show simulated Markov chains with $\sigma_x = 0.5$ and $\sigma_0 = 1.0$. The solid line in each graph is a sequence of sampled values of x_n, and the dotted line and the gray area show the mean and 95% confidence interval on x_n. All samples and statistics are conditioned on $x_1 = 20$. Panels (d)-(f) show a histogram of the last 80 values of x_n for the Markov chains in panels (a)-(c). The gray areas are the probability density functions for the stationary distributions of the three models. 49
4.2 Initial and final distributions, and the chains for the two conditions in Experiment 1. (a) The histograms for the initial distributions of stimuli for conditions A and B. (b) The chains of serial reproduction, starting from the first participant's data, ending with the 18th participant's data. (c) The histograms for the ending distributions of stimuli for conditions A and B. 53
4.3 Autoregression and bias plots for the two conditions in Experiment 1. (a) The autoregression plots for conditions A (black) and B (gray). (b) The bias plots for conditions A (black) and B (gray). 54
4.4 Stimuli, training distributions and stationary distributions for Experiment 2. Each data point in the right panel shows the mean of the last 20 iterations for a single participant. Boxes show the 95% confidence interval around the mean for each condition. 55
4.5 Chains and stationary distributions for individual participants from the two conditions in Experiment 2. (a) The three Markov chains generated by each participant, starting from three different values. (b) Training distributions for each condition. (c) Gaussian fits for the last 20 iterations of each participant's data. (d) Autoregression for the last 20 iterations of each participant's data. 56
4.6 Autoregression and bias plots for the two conditions in Experiment 2. (a) The autoregression plots for conditions A (black) and B (gray). (b) The bias plots for conditions A (black) and B (gray). 57

4.7 Chains and stationary distributions for individual participants from the two conditions in Experiment 4. (a) The four Markov chains generated by each participant, starting from four different values. In condition A, black and white plot markers are the red and blue fish chains respectively, and in condition B, all the chains are gray fish. (b) Training distributions for each condition, black and light-gray plots being the red and blue fish respectively. (c) Gaussian fits for the last 20 iterations of each participant's data. Black and light gray plots in condition A are the red and blue fish chains, and both plots in condition B are the gray chains. 62

5.1 Simulating the cultural transmission of color term systems. (a) The array of colors used in the World Color Survey (WCS). Members of 110 non-industrial societies reported the terms that their languages used to label these colors. (b) Examples of color-term systems produced by simulating cultural transmission in the laboratory. Each column shows one chain of systems produced by participants learning novel color terms from examples sampled from the system of labels assigned to colors by the previous participant. Transmission proceeds down the column, and different columns show chains for systems with two, three, four, five, and six terms. The color chips are arranged in the same order as in (a), so the position of a chip in the arrays corresponds to its color. The colors signify which chips were labeled with the same color word, and do not directly correspond to the colors denoted by the words (light blue is used to indicate a minority term, used for fewer than 5% of all chips). The first system in each column is a randomly generated initial partition. The last system in the first column is the Dani language, for which only aggregate data using a subset of the WCS array are available (the gray bars correspond to unlabelled chips). The last five systems in the remaining columns are data from individual speakers of the closest matching WCS language (determined by averaging Variation of Information (VI) values across iterations 4-13 of each chain). 70

5.2 Comparing systems of color terms produced by simulating cultural transmission in the laboratory with those from non-industrial societies, as represented by the World Color Survey (WCS) and English color systems. The distance between systems is assessed using Variation of Information (VI), an information-theoretic measure of the difference between two partitions of the same set of elements. (a) The mean VI between systems produced by English-speaking participants and those seen in the WCS decreases as the number of iterations of cultural transmission increases. The VI converges to a level that is not statistically significantly different from the mean VI between languages in the WCS. (b) Taking the final systems produced by our laboratory participants, rotation along the hue dimension of the WCS array results in a worse fit to the WCS, confirming that there are significant similarities between the way the languages emerging in our experiments, and those spoken in non- industrial societies, categorize colors in terms of the hue dimension in the WCS. The horizontal axis indicates the number of steps the chips were rotated to the right along the hue dimension of the array shown in Figure 1. (c) The mean VI between systems produced in the iterated-learning experiment and those produced by participants applying English color terms. The VI between systems in iterated learning chains and English systems did not reduce to the same level as the mean VI between systems produced by English speakers. 75
5.3 Mean distance from the WCS systems for chains with different numbers of terms and different initial partitions. 77
5.4 Mean VI from the WCS systems for chains with varying numbers of terms and different initial partitions, when comparisons were only made with WCS languages that use an equal number of terms. 78
5.5 Comparing systems of color terms produced by iterated-learning with WCS and English color systems. The distance between systems is assessed using the Variation of Information (VI) in the first row, the Adjusted Rand Index (AR) in the second row, and the Van Dongen metric (D) in the third row. The first column shows the mean distances between systems produced by English-speaking participants in our iterated learning experiment and those seen in the WCS. The second column shows the final systems produced by our laboratory participants rotated along the hue dimension of the WCS array, resulting in a worse fit to the WCS using all three metrics. The third column shows the mean distances between systems produced in the iterated learning experiment and those produced in the English experiment. 80

5.6 Examining the behavior of the proportion of matching pairs n_{11} and mismatching pairs n_{00}. The first column shows the mean distances between systems produced by English-speaking participants in our iterated learning experiment and those seen in the WCS. The second column shows the final systems produced by our laboratory participants rotated along the hue dimension of the WCS array. The third column shows the mean distances between systems produced in the iterated learning experiment and those produced in the English experiment. 82

List of Tables

2.1 Behavioral results for experiment 1a and 1b. Mean (SE) RT for go and failed stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop. 17
2.2 Mean (SE) amplitude (in millivolts) of MEPs recorded from each muscle in experiment 1a (N = 14). 17
2.3 Behavioral results for Experiment 2 (N = 10). Mean (SE) hand and foot RT for go and failed stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop. 21

3.1 Behavioral results for Experiment 1 (within-hand). Mean (SE) RT for Go and Failed Stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop. 34
3.2 Mean (SE) amplitude (in millivolts) of MEPs recorded from each muscle in before each type of behavior in Experiment 1. 35
3.3 Behavioral results for Experiment 2 (between-hand). Mean (SE) RT for Go and Failed Stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop. 37
3.4 Mean (SE) amplitude (in millivolts) of MEPs recorded from each muscle in before each type of behavior in Experiment 2 37
3.5 Behavioral results for Experiment 3. Mean (SE) RT for Go and Failed Stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop. 39
3.6 Mean (SE) amplitude (in millivolts) of MEPs recorded from each muscle in before each type of behavior in Experiment 3. 41

Acknowledgments

One individual stands out and speaks up, but her voice is the voice of all human kind. The successful story in science is the story of a joint effort for the common benefit. Here I am, privileged and humbled to be the person to present what we have found within the science of the mind. I'm obliged to thank everyone who has helped to hold the pillars of this stage.

Richard B. Ivry, my intellectual father, has been constantly engaging me in scientific dialogues, challenging me to think critically, encouraging me to explore unknown territories, and inspiring me to seek a vision in science and in life. His detailed support has not only pushed me over countless technical and theoretical hurdles in the first line of research in this dissertation, but has guided me though a path of self-discovery. His incredible amount of knowledge in the field of cognitive neuroscience, razor sharpness in dissecting a topic, amazing ability of pulling together seemingly un-related topics and generating new insightful observations to answer fundamental questions in cognition, as well as his perspectives and engagement in every aspect of life are a constant inspiration. Thank you, Rich, for being an amazing adviser who always has others' best interest at heart, a great role model and a lifelong friend.

Thomas L. Griffiths, my co-intellectual-father, has a brilliant mind and a vision in science. I am most grateful for his mind-changing teaching on the philosophy and history of science, numerous topics in cognitive science, and instilling in me a probabilistic and mathematical view of the world. His hands-on guidance on mathematical and computational skills has not only guided me through accomplishing the second line of research in this dissertation, but has also brought me to a broader understanding of cognitive science. His humbleness, his respect for others, and his self-discipline have been and will continue to be a great model for me to follow.

I could not appreciate more for being a member in such a wonderful and open community as Berkeley. It is the place I consider as home. It is here that I learned how to love and how to trust. The following is only an incomplete list of those who have offered a selfless hand here and there along my graduate years.

Terry Regier, the outside member of my dissertation committee. I would like to thank him for insightful comments and ideas to my research on color categorization, and for his teaching on how to handle the huge amount of data collected in the World Color Survey. Thanks to Marty Banks and Jonathan Wallis, my qualifying exam committee members, for passing along your wisdom without reservation each time we discussed my research questions.

Thanks to Julie Duque and Tim Verstynen, who have given me hands-on instructions on the TMS technique, and to Adam Sanborn, who assisted me in setting up iterated-learning experiments. To Andrea Weinstein and Neil Albert who have taught me a great amount of skill in testing patients with Parkinson's disease. I am most grateful for all of your guidance.

I would like to thank my collaborators, Mike Dowman, and Zach Westrick, for your hard work and commitment, your ideas and discussion on the color and stop-signal projects. Geoffrey Brookshire also helped in collecting data for the stop-signal project. Thanks to Tony Lai, Linsey Smith, Jason Martin, Brian Christian, Vladimir Shut, Aaron Beppu, Ryan Waliany, Katherine Kolh, and Matt Cammann for their assistance with collecting the large amount of data in the iterated-learning projects. It's been a great pleasure working with each of you, and enjoyed your lively discussions on various topics.

Any successful project has a great support team behind the scene. As for me, I've been lucky enough to have more than one. I would like to thank my two wonderful labs, the members of Cognition and Action (cognac) and Computational Cognitive Science (cocosci), for sharing all the beautiful days of graduate school, for your stimulating discussions, teachings on everything about the Western culture, and all the fun times together that have enriched my life and extended my horizons. I had a fun time collaborating with John Schlerf, Nola Klemfuss, Becca Stoloff, Jordan Taylor, Arne Ridderikhoff, Jessica Mollick, Robert Musgrave, Shawn Ell, and Peter Butcher on several other projects that are not presented here. Special thanks to Kevin Canini, Chris Lucas, Saiwing Yeung, Laura Hieber, and Dav Clark for various technical supports throughout these years.

I would like to thank my friends scattered around the world, Yun Ji, Yingshi Gao, Xiaowei Lu, Ludovica Labruna, Yula Paluy, Saffron Walkling, to name a few. Each of you, time and again inject me with your incredibly fresh ideas and stun me with your beautiful, balanced view of life. And thanks to Weiguo Zhang for being a special, understanding and supporting friend, for sharing and exploring together in the past 10 years.

Lastly, I would like to thank my parents, Huishan Sun and Siyuan Xu, my brother Ming Guo, and others in my big family back in China, for their constant support and unconditional love.

Chapter 1

Introduction

People constantly strive to maintain a stable position in the world, to feel balanced. However, change, and therefore imbalance, is inevitable. Thus being in control is an illusion. How we maintain this balance and how we adapt to the changes are two sides of the same coin, essential questions in cognition and in human existence. What does it mean to be in control, yet adaptive? A big part of the effort involves allocating resources to follow or fight against the various flows of influences, inferring and updating our models of the world, and constantly guiding the execution of actions. Structure and coherency is what we seek among the seemingly chaotic objective reality. Along the road, everything is filtered through our own mind, our mental representations of the world and our attention and awareness at a particular moment.

Here I attempt to present several empirical observations and interpretations, collected along the road from the self to the world, in an effort to understand a structured and coherent answer to these questions through the lenses of science.

1.1 Flexibility and biases in cognitive control: Hardware or software?

Adaptation demands that we can constantly modify our behavior. Even with the best-planned action, things can happen unexpectedly. You reach for a bottle of water, and forget that it's already empty; you step on an apparently sturdy rock and realize that it's quite slippery after a quick shower of rain. Flexibility lends us the ability of mapping out our new surroundings, modifying our representations, and tuning our bodies to quickly adjust to the new situation.

In the first line of research, I examine flexibility and its limitations from the angle of inhibitory control. Specifically, in Chapter 2, I ask if people can selectively stop some part of an initiated action without disrupting other parts of that action. This line of research follows Logan's seminal model of a horse-race between going and stopping (Logan & Cowan, 1984), and a literature of inhibitory control using the stop-signal paradigm (Lappin & Eriksen, 1966; Logan, Cowan, & Davis, 1984; Logan, 1994; reviewed in Verbruggen & Logan, 2008). Previous studies have

shown that people are unable to selectively stop, hallmarked by what is referred to as a Selective Stop cost, the delay in reaction time for the action that is not aborted. This limitation has been attributed to hardware constraints arising from a brain structure, the subthalamic nucleus (STN), the presumed source of stop signals (Aron & Poldrack, 2006). We used single pulse transcranial magnetic stimulation (TMS) as a probe on the excitability of motor pathways, examining the specificity of inhibition when participants perform a selective stop-signal task. We also trained people to selectively stop in a second study, testing a Sampling-Bias Model, within the framework of the horse-race model. This model provides a novel account of the Selective Stop cost. As a package, this study shows that people can selectively inhibit our actions.

The question is then how is flexible control achieved? Looking for a "controller" will inevitably drag us down the rabbit hole of infinite regress. Constraint seems to be the key to answering these questions. Our empirical findings and modeling results point in the direction of a more plastic view of the constraints that limit people's ability to selective stopping, similar to many other cognitive constraints, such as those exhibit in multi-tasking situations (Broadbent, 1958). We show that these constraints may share a similar mechanism as shown in other conflict resolving situations, such as in the flanker and Simon tasks and dual-tasks (Ridderinkhof, 2002), and they can be overcome. As such, the ability to selectively stop can be learned.

One important factor that contributed to learning was the use of a salient stopping cue. This finding suggests one way that flexibility can be built on top of constraints. In addition, it was also important to establish the appropriate utility function, an incentive in the task that rewarded obeying the task instructions. A basic observation here is that flexibility and biases are not in opposition to one another. They operate hand-in-hand. It is simply impossible to be unbiased, and control cannot be achieved without constraints. Flexibility allows us to take advantage of these biases to achieve cognitive goals. (An amusing side but not completely irrelevant note is that it is the insight of the statistical sampling bias in the stop-signal paradigm that lends us the hammer to nail the answer to the ground.) Understanding these flexibilities and biases is the same thing of knowing our limitations, which is liberating.

The remaining questions are then where the biases come from, how they are built in the system, and what do they offer us. One way to view the constraints is biological, such as reflexes and patterns of muscle co-contractions. This view, however, is not completely satisfying. Even if we have a complete inventory of these muscle contractions, do we completely understand the controlling processes so as to build a machine that can control properly? For example, how do we handle the millions of possible combinations and permutations of our muscles so that they successfully allow us to push a thread through a needle hole? IBM's Deep Blue would be dumbfounded if facing a task that a five-year old girl may easily achieve. We need other tools and some structures to talk about these patterns. The horse-race model, for example, offered one of the simplest high-level views of the underlying control-processes for inhibition. This model abstracts away the details of processes involved in initiating and aborting actions, but it offers us a handle on an interesting question. From motor learning and motor control perspective, these models capturing underlying processes and patterns are internal models and maps that we are constructing and fine-tuning in our mind throughout our lifetime experience. In the following sections, we will focus on these ques-

tions, and hopefully convince the reader that those motor maps may not be theoretically different from other kinds of cognitive biases, often called categories, and thus may be understood using similar approaches.

1.2 Surfing the waves of fluctuations in cognitive control: How do unknown biases influence us?

The common mystery and self-centered mistake in cognition is to think that control is voluntary. It gives us the illusion of being in control. When this illusion fades away, we feel the limitations of our own control, or even feel being controlled. Forces come from all sorts of tendencies, habits, reflexes you name it. We can call them unknown biases. The fluctuations might be roughly classified into two different categories: idiosyncratic traits or internal states.

The brain is often viewed as an information processing system, which takes input from the external world and emits some output after some form of internal processing (Marr, 1982). The majority of psychology and neuroscience studies focus on mapping a function between the input and output so as to infer internal states. Internal states that appear to be independent from the external world are often too hard to control and are thus avoided, or at the best, regressed away as the "noise." But they have not been completely neglected. James (1890) advocated the method of introspection. More advanced technologies and computational tools now offer some insight in those long avoided topics in cognitive science, for example, a more recent endeavor in fMRI studies is to study the functional role of the "resting-state" (reviewed in Fox & Raichle, 2007). Sejnowski (1981) and other researchers (e.g., Malsburge, 1994) suggest another approach, viewing internal fluctuations as dynamics that may carry some functional meaning.

In Chapter 3, I report three stop-signal experiments that look into the functional role of those fluctuations in inhibitory control. We used the single-pulse TMS to probe the baseline excitability in people's motor pathways, asking if the fluctuations in these signals predict the behavioral outcome of stopping. The findings are mixed. The modulation of the intrinsic fluctuation was only observed in a within-hand version of the stop-signal task and was absent in a between-hand task. More puzzling, we failed to replicate these results in a subsequent study. However, a sequential analysis shed light on the interaction of stimulus-driven signals and these intrinsic fluctuations. We found that when the inhibition was successful during the previous trial, a signature of this control signal carried over to the next trial. Moreover, this carry-over effect was task-specific, limited to the selected muscle. In contrast, the heightened state of the fluctuation during a Failed Stop could be either ramped down quickly or remained high, depending on the effectiveness of inhibition.

It remains unknown if the fluctuations are causes or effects. What are the rhythms and what are those unknown biases? Salinas and Sejnowski (2001) suggest that a coherent picture of neural synchronization might be a biasing influence controlling the strength of signal from up-stream circuits to down-stream circuits. In the second line of research, I present a higher level view of the information flow, a probabilistic view.

1.3 Positioning ourselves in a probabilistic world: Empirical and computational approaches

Much of the effort in scientific discoveries involves causal inference, searching for the function that maps the input to the output. However, we, often end up with correlations. Is causal inference possible at all? Perhaps. Hume's (1748) assumption that tomorrow is the same as today is a rather liberating notion, which, paradoxically, gives us a balanced and unbiased view of the world at a larger scale. It allows us to go about collecting data and testing our hypothesis, checking and updating our representations of the world.

This notion, however, only partially resolves our philosophical uncertainty. In daily lives, uncertainty exists everywhere. This is when the probabilistic view comes into the picture. Things in the temporal domain can be viewed as stochastic events, while representations and maps are distributions of functions. We capture the underlying patterns of these events and distributions using formal statistical models rather than traditional linear functions. These probabilistic lenses offer us much richer information and more powerful computational tools.

As revealed in the first line of research above, the simple Selective Stop cost measure and traditional hypothesis testing approach may not be sufficiently sensitive to pick up critical differences. Also, as physiologists mapping out an inventory of all the constraints in our body, the huge amount of information often seems intractable. An additional problem comes from multiple-comparisons that arise in hypothesis testing (Cournot, 1843), a problem that is more prominent in the second study on the effect of intrinsic fluctuation. While powerful new technologies, such as fMRI, open exciting new doors to research, they also introduce a large amount of uncertainty (Vul, Harris, Winkielman, & Pashler, 2009, but also see Lindquist & Gelman, 2009). The multiple-adjustment problem is nearly inevitable. All these are calling upon more sophisticated statistical tools, such as time-series analysis and sampling methods.

Back to causal inference, the empirical approach allows us to intervene and therefore find causal relationships among variables. Pearl (2000) provided us a powerful approach combining the traditional logical approach and the probabilistic view: graphical models. In the second line of the research presented here, we use one simple graphical model, the Markov chain. Specifically, we view information transmitted from one agent to another as a Markov chain (see section 1.4). Assuming certain statistical properties of the nodes along the chain, we can gain some surprising insight concerning the dynamics of the information flow.

A coherent approach for understanding human learning and inference can be achieved using Bayesian inference, a framework which views the world as degrees of beliefs of events. With this approach, we assume people use Bayes' Theorem to update their beliefs about the world. If we assume the agents transmitting information along a chain are doing Bayesian inference, an interesting leap can be achieved: We can analyze information flow from one agent to another as a Markov chain and this process is equivalent to a sampling method called Markov Chain Monte Carlo that is used to approximate complex distributions in statistics. This insightful connection gives us a powerful tool to grasp seemingly intangible biases in the human mind. The second

line of research examines this way of exploring the information flow: Discovering our mental representations of the world using iterated-learning procedures.

1.4 Representing the world: what are the biases in our mind and in the world out there?

One way to view the rhythm of the world is to consider how information is transmitted from one part of the body to another, one person or one culture to another. How can we capture these dynamics as a way to understand an entity? One elegant model, alluded to in the previous section, is iterated learning, which assumes that each generation of learners receives information from the previous one, generates some more information, and passes it on to the next generation (Kirby, 2001). This method has been used in anthropological and social psychology studies to simulate cultural evolution (reviewed in Mesoudi, 2007; Mesoudi, Whiten, & Laland, 2006). Griffiths and Kalish (2005, 2007) formally analyzed the process of iterated learning as a Markov chain and showed that if the learners are assumed to be Bayesian and sharing common inductive biases, as represented by a prior distribution, this process will converge to the stationary distribution that reflects people's inductive biases. In two studies, we attempt to use this model to discover the biases that might influence categorization.

In Chapter 4, I present a study in which that we apply the iterated learning model to reproduction from memory and explore the biases in serial reproduction. Serial reproduction is a paradigm developed by Bartlett (1932) in which one person's reconstruction from memory became the stimulus seen by the next participant. Bartlett concluded that this process magnifies cultural biases. We formally analyze serial production using a Bayesian model and show that this process converges to the prior distribution reflecting people's inductive biases. We also test the predictions of this model in laboratory using four serial reproduction experiments with strictly controlled stimuli. We show that iterated learning can be use to investigate the inductive biases of human learners.

Our findings in the first study lend us a powerful tool to move onto discovering other biases, ones that might underlie cultural universals. In the second study (Chapter 5), we used an iterated-learning paradigm to discover linguistic universals in color naming. We simulated the transmission of color words, asking English-speaking participants to learn artificial color term systems. We show that the color-term systems generated in our laboratory experiments converged to be consistent with those seen across the 110 languages of the World Color Survey (Kay, Berlin, & Merrifield, 1991; Kay, Berlin, Maffi, & Merrifield, 1997; Kay, Berlin, Maffi, Merrifield, & Cook, 2009). This finding suggests that some of cultural universals could be explained in terms of the dynamics of cultural transmission bringing out our perceptual and learning biases.

Iterated learning provides a unique approach for identifying other seemingly intangible biases, and also offers an opportunity to capture the dynamics of information flow. Using the basic structure of Markov chain, we can explore the dynamics of many different kinds of information flow. For example, if we add a node for data collection at each temporal step along the chain, this be-

comes the classical state-space model (Wiener, 1949), and within this framework, adaptation can be viewed as a constant process of collecting data from the world, using them to evaluate models in our hypotheses space, and passing them along to the next time point within our own mind, or to the generation. Under this umbrella, a rich set of statistical models and learning methods can be applied to answer some interesting questions in control and adaptation. Particularly, sampling methods extends our horizon from pre-assumed distributions such as linear dynamics with Gaussian noise to any shapes.

As a final note, once we cross the boundary of individual and culture, internal or external, innate or learned, we see the importance of understanding our limitations and finding the rhythm of the world. Hume's assumption may or may not be true, but as long as we follow the pattern of Markov chain, we may capture the mental representations in the human mind. Moreover, as Beppu and Griffiths (2009) showed that if we allow learners observe actual data from the world and also to pass along the actual beliefs on to the next generation, the process of iterated-learning converges to the hypothesis that captures the distribution closest to the state of the world. This analysis gives us the hope that although each single agent is biased and may not be fully informed the common effort may eventually converge to the truth.

Chapter 2

Flexibility and Specificity in Inhibitory Control

2.1 Introduction

Flexibility is an essential feature of cognitive control. When talking with a friend, we readily modify a planned utterance when another individual joins the conversation. In shooting a basketball, a skilled player can modify a planned shot to avoid a defender. Flexibility frequently requires the inhibition of one action in favor of another. In some situations, only part of the planned action must be changed. When talking to a friend while walking in the city, the appearance of a fast-moving car may require that we abruptly abort a plan to cross the street. This unexpected stimulus, however, need not require that we temporarily abort our conversation. We ask here whether people are capable of selective inhibition.

The *stop-signal task* has been extensively used to study inhibitory control (Lappin & Erikson, 1966; Logan & Cowan, 1984; Logan et al., 1984; Logan, 1994; Verbruggen & Logan, 2008). In this task, participants are instructed to make a speeded response. On some percentage of the trials, a stop-signal dictates that the participant attempt to abort the planned action. The likelihood of stopping has been modeled as a *horse race* (Logan & Cowan, 1984) between signals associated with initiating the response ("Go") and signals associated with aborting the response ("Stop"). At the neural level, inhibitory signals have been linked to cortico-basal ganglia interactions (Mink, 1996), with an emphasis on the hyperdirect pathway involving inferior frontal gyrus, supplementary motor area, and the subthalamic nucleus (STN) (Aron & Poldrack, 2006; Aron et al., 2007). These regions show an increase in activity on stop trials, with the magnitude of this signal greater on trials in which the participants were successful in stopping (Aron, 2006, but see Xue et al., 2008).

The studies reviewed above have generally involved stop-signal tasks entailing a single response. It is unclear if similar processes and neural mechanisms are engaged during tasks in which only one component of a planned action must be inhibited. Based on an event-related potential study, De Jong et al. (1995) proposed that distinct forms of inhibition operate in the standard and selective stop signal tasks. For the former, a subcortical mechanism operates in a generic manner, clamping down the peripheral motor system in the face of normal cortical

preparatory signals. For the latter, a cortical mechanism can be invoked to selectively inhibit a specific action. Behaviorally, De Jong et al. observed a cost in selective stopping, with an increase in reaction time for the non-inhibited response on selective stop trials compared to go trials.

This delay was also observed in a series of studies by Coxon et al. (2007), leading these authors to suggest that selective stop actually involves two processes. First, a generic inhibitory signal is generated that freezes all planned actions, perhaps reflecting the poor topographical resolution of the STN. Following this, the non-inhibited response must be rapidly reprogrammed. Thus, the selective stop cost reflects a structural constraint imposed by the initial engagement of a generic stop command that is directed at both the inhibited and executed components of the planned action.

We report here two experiments to test this hypothesis. In Experiment 1, we use transcranial magnetic stimulation (TMS) to examine changes in corticospinal (CS) excitability in task-relevant and task-irrelevant effectors following the onset of a stop signal. In Experiment 2, we use a model-based approach to re-examine the behavioral costs observed in selective stop tasks. The results of both studies challenge the notion of generic inhibitory signals. Rather, the results point to a softer constraint that reflects the degree of overlap between representations of the selected and non-selected components of a multi-component action.

2.2 Materials and Methods

2.2.1 Participants

A total of 40 right-handed healthy participants participated in Experiments 1 and 2. Handedness was confirmed via a condensed version of the Edinburgh Handedness Inventory (Oldfield, 1971). All participants gave written informed consent and were financially compensated for their participation. The protocol was approved by the institutional review board of the University of California, Berkeley.

2.2.2 Experiment 1

Transcranial Magnetic Stimulation.

In Experiment 1, corticospinal (CS) excitability was assessed by measuring motor-evoked potentials (MEPs) elicited by single-pulse TMS applied over the left or right primary motor cortex (M1). The TMS pulses were applied with a figure-of-eight coil (diameter of wings 70 mm) connected to a Magstim 200 magnetic stimulator (Magstim, UK). Participants wore a tight-fitting electroencephalography (EEG) cap and the coil was placed tangentially on the scalp, with the handle oriented toward the back of the head and laterally at a 45º angle from the midsaggital

line. The position of the stimulator on the skull was adjusted so as to produce a maximal response from the first dorsal interosseous muscle (FDI). At this location, responses are also observed in the abductor digiti minimi muscle (ADM). The location was marked on the EEG cap to provide a reference for the experimental session.

The resting motor threshold (rMT) was defined as the minimal TMS intensity required to evoke MEPs of ~50 μV (peak-to-peak amplitude) in the targeted muscle on five out of 10 consecutive trials. Thresholds were defined over both left and right motor cortex in Experiment 1a, and corresponded to 50.1% (SE = 2.0) and 49.4 (SE = 2.2)of maximum stimulator output (MSO), respectively. Only left motor cortex was stimulated in Experiment 1b and here the mean stimulation level was 48.3% (SE = 1.7) of MSO. The intensity of TMS during the experimental session was always 115% of rMT, set on an individual basis.

Electromyography (EMG) recording.

EMG activity was recorded from surface electrodes placed over the FDI and ADM muscles of both the left and right hands. We also recorded EMG from a leg muscle, the right soleus, in Experiment 1b. EMG data were collected for 3 sec during each trial, beginning 150 ms before the onset of the fixation marker (see below). The EMG signals were amplified and bandpass filtered on-line 50-2000 (Delsys). The EMG signals were used to measure peak-to-peak amplitudes of the MEPs. Trials with background EMG activity > 20 μV in the 200 ms window preceding the TMS pulse were excluded from the analysis (Duque et al., 2005, 2007; Duque and Ivry, 2009).

Experimental Task.

Changes in corticospinal excitability were monitored while participants performed a stop-signal task. In Experiment 1a, participants (11 women, 3 men; 22.1 ± 2.6 years old) performed a two-choice RT task, activating a response key with either abduction movement of the right index finger (FDI) or flexion movement of the right pinky finger (ADM). The task-irrelevant, left hand rested on a pillow during the entire experiment.

Each trial began with a 500 ms presentation of a fixation marker, an asterisk, at the center of the screen (Figure 2.1A). This was followed by a green circle, the "Go" signal, that appeared on either the left or right side of the computer screen (10 deg eccentricity). Participants were instructed to make a speeded response, pressing on a response device with the spatially compatible effector (e.g., index finger when stimulus was on the left side). On a fixed percentage of the trials (33% during training and 40% during the TMS sessions), the green color changed to red, the "Stop" signal. On these trials, participants were instructed to try and abort the planned response. The interval between the onset of the green circle and the color change on stop trials is referred to as the stop-signal delay (SSD), determined in an adaptive manner (see below). The circle remained visible for 1000 ms after the SSD and the screen then went blank for 2500 ms (inter-trial interval).

An adaptive, staircase procedure was employed to determine the SSD on a trial-by-trial basis (Osman et al., 1986, 1990; Band et al., 2003). The SSD value was increased by 50 ms after a successful stop and decreased by 50 ms after a failed stop. This method converges to a SSD value at which participants will successfully stop a planned response on approximately 50% of the trials (Levitt, 1971). Separate staircases were run for FDI trials and ADM trials, given that the reaction time for these two muscles may not be equivalent. Moreover, for each muscle, there were two separate staircases, one starting at 50 ms and the other at 300 ms. Participants were informed of the general nature of the adaptive procedure and that they should expect to fail to stop on a large number of trials. They were also instructed that the procedure minimized the effectiveness of adopting a strategy to purposefully slow down in order to check for the color change.

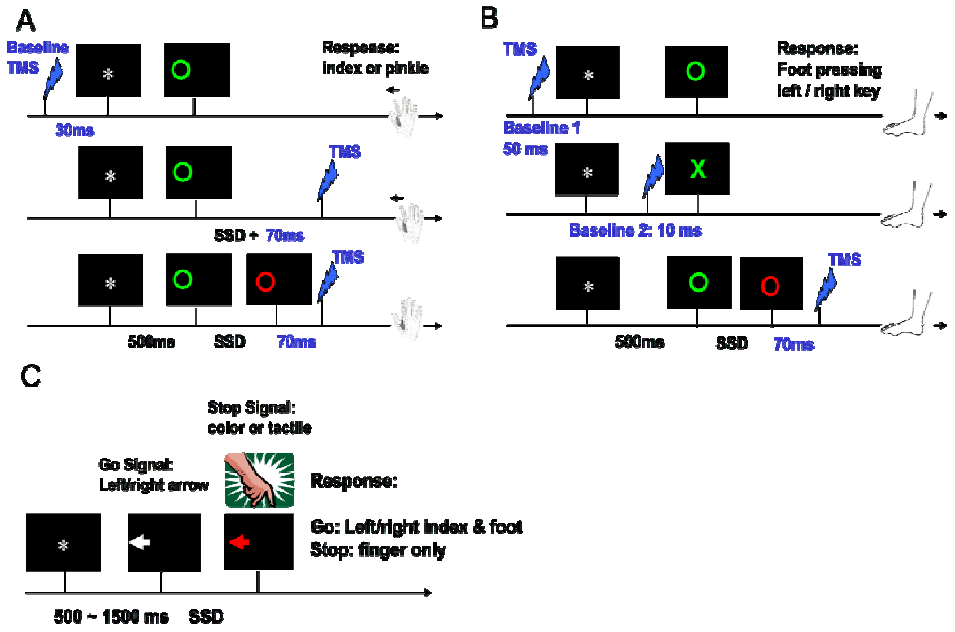

Figure 2.1. Stop-signal task procedure and TMS timeline used in Experiment 1a (A), and Experiment 1b (B). C. Stop-signal task procedure used in Experiment 2.

Each participant completed two sessions on two separate days, one with TMS over the right M1 and one with TMS over left M1. The order of sessions was counterbalanced across participants. Each session began with three practice blocks of 72 trials each (48 Go trials, 24 Stop trials). The SSD estimates at the end of each block were used as the initial values for the subsequent block. The training blocks were followed by four TMS blocks. The mean SSDs in the last training block, one for the index finger and one for the pinky finger, were used as a fixed SSDs in the TMS blocks. We opted to keep the SSD fixed so that the timing of the TMS pulse remained constant with respect to the onset of the Go signal. To prevent participants from adjusting to this fixed SSD, eight "catch" trials were included, four in which the SSD was

reduced by 80 ms and four in which the SSD was increased by 80 ms. Thus, each TMS block consisted of 80 trials, 48 Go and 32 Stop.

One TMS pulse was applied on each trial, occurring either 30 ms before the onset of the fixation marker (baseline TMS) or 70 ms after the SSD. The 70 ms timing was selected based on pilot testing used to determine a value at which the TMS pulse would arrive after the onset of the color change on Stop trials while falling within a window of 120 – 20 ms before EMG onset on Go trials. To maximize TMS probes after Stop signals, the baseline pulses were only obtained on 1/3 of the Go trials (participants were not aware of this). The post-SSD pulses occurred on the other 2/3 of the Go trials and 100% of the Stop trials. Note that the timing was the same for both Go and Stop trials, even though the color change only occurred on Stop trials.

Experiment 1b (11 women, 5 men; 22.6 ± 3.4 years old) provided a second assessment of changes in corticospinal excitability in task-irrelevant muscles. The procedures were the same as in Experiment 1a with three exceptions. First, the Go signal was the letter X or O, appearing at the center of the screen (Figure 2.1B). Participants used their right foot to press either a left or right button on a foot-designed response board, with the correct button based on the identity of the letter. We used a non-spatial choice task to minimize any possible effect of spatially-displaced stimuli. Second, TMS pulses were only applied over right M1, with the coil positioned to optimally elicit MEPs in FDI of the left hand. Activation of a right foot response should have minimal opportunity to directly influence excitability changes in the left hand. Third, we added a second baseline TMS time, 10 ms prior to the onset of the Go signal (in addition to the one appearing 30 ms prior to the onset of fixation). This second baseline allowed us to assess changes in excitability that may occur in anticipation of the trial (Sohn, et al., 2003).

Behavioral data.

We used the reaction time (RT) and stop-rate data to assess participant's performance in the stop-signal task. RTs were recorded by the activation of microswitches on the response board. We also estimated the stop-signal reaction time (SSRT) according to the horse-race model, which is defined as the median of the Go RT distribution minus the mean of the SSD (Logan & Cowan, 1984; Band et al., 2003).

MEP data accquisition and analysis.

MEPs were identified using custom Matlab software (MathWorks) from the EMG recordings in the (-200, 1000) ms time window around the TMS pulse. The MEPs can be sorted into four categories: Baseline, Go, Successful Stop, and Failed Stop To directly compare MEPs on Go and Failed Stop trials, we only included data for these conditions from trials in which the TMS pulse occurred 120 to 20 ms before EMG onset. As a result, for eight out of the 14 participants, we did not have enough number (≥ 8) of trials for calculating average MEPs for Failed Stop. For this reason, we won't discuss the Failed Stop results in details in the next section. EMG onset was manually identified with the criterion of peak-to-peak waveforms > 50 μV and lasting longer than 50 ms. These EMG onset RTs were then checked against RTs recorded from the

response board to confirm the timing. We cannot apply this rule on Successful Stop trials since there is no movement; thus, we included all of the trials in this category.

We also eliminated trials in which there was evidence of an increase in background EMG activity (all categories) (12% trials), Successful Stop trials with partial EMG activity around the TMS pulse (using the same above window and criterion) in the selected muscle (7% trials), Go or Failed Stop trials with partial EMG activity in the non-selected muscle, as well as trials with EMG evidence of mirror movements in a muscle of the resting hand homologous to the response agonist in a window ± 50 ms of EMG onset (5% trials).

Our primary interest is in the modulation of the MEPs as people perform the stop-signal task, comparing how these changes are manifest in both selected and non-selected muscles. We classified the muscles based on their functional role. In Experiment 1a, a muscle was *task-relevant* when it was part of the response set (FDI and ADM of the responding hand). Within this category, a muscle was a *selected* respondent when the cue indicated that the muscle would be the agonist for the forthcoming response, and *non-selected* otherwise. Muscles that were not part of the response set were classified as *task-irrelevant, and could either be homologous* or *non-homologous*, based on their relationship to the selected respondent. In Experiment 1b, both of the probed muscles (left FDI and ADM) were task-irrelevant and non-homologous.

For each of these categories, we evaluated the MEPs, comparing the changes across the trial, relative to baseline. We also compared the MEPs on Successful Stop trials to those obtained on Go trials.

2.2.3 Experiment 2

Experimental procedure.

In experiment 2, 10 participants (7 women, 3 men; 23.2 ± 5.8 years old) were tested on a selective stop-signal task (Fig. 1C) to assess the cost associated with aborting one response of a multi-component movement. Because we were interested in whether a selective cost is inevitable due to the recruitment of a generic stopping process, we designed the experiment to maximize the opportunity for participants to avoid this cost. This was achieved by providing the participants with multiple days of practice and monetary rewards, as well as by assessing the benefit of a highly-compatible stop signal.

The trial began with the participants resting their index fingers on two response keys and the right foot on a foot pedal. An asterisk was presented for a random delay (500 ms to 1500 ms) to serve as a fixation and alerting marker. After the delay, the asterisk was replaced by a Go signal, an arrow pointing to the left or right. This signal indicated two responses, a manual choice response and an invariant foot response. The manual choice was between the left or right index finger, with the selected finger corresponding to the direction of the arrow. For the invariant response, the participants pressed a foot pedal with the right foot. The arrow disappeared when a response was detected, or after 2 sec on successful stop trials.

On 33% of the trials, a stop signal indicated that the participant was to selectively abort the manual response. As in Experiment 1, two interleaved staircases, one starting at 50 ms and the other at 300 ms, were used to determine the stop signal delay (SSD). In separate sessions, we compared two different stop signals. In the visual condition, the color of the arrow turned red. In a tactile condition, an A-to-D signal from computer caused the response key associated with the cued finger to vibrate. We reasoned that the high degree of compatibility here would enable participants to readily identify the response that should be aborted, eliminating any cost associated with mentally assigning a relatively abstract stop signal (e.g., color red) to one of two prepared responses.

The instructions emphasized that the participants should make the two responses as quickly as possible, responding simultaneously with the finger and foot. On Stop trials, only the manual response was to be aborted; the foot response should be made without hesitation. As in Experiment 1, the instructions emphasized that it would not be always possible to abort the manual response on Stop trials and that the participants should avoid adopting a strategy of slowing down to increase the likelihood of successfully stopping.

Participants were trained on four different days, with each type of stop signal (visual or tactile) used on two consecutive days. The order of conditions was counter balanced across participants. During each day, they first performed two pure "Go" blocks of 32 trials each, containing no stop signals. Following this, they completed 10 selective-stop blocks, each composed of 60 trials (40 Go, 20 Selective Stop).

To provide an incentive for participants to avoid delaying the foot response on subsequent selective stop trials, the mean foot RT from the second of these pure Go blocks was used to establish a benchmark for determining subsequent monetary bonuses. At the end of each of the 10 selective-stop blocks, the mean foot RT was computed and displayed to the participant, along with the amount of money earned on that block based on the following criteria: 1) $1.00 if the mean foot RT was less than the standard; 2) $0.75, $0.50, and $0.25 if the RT was within 25 ms, 50 ms, and 75 ms, respectively of the standard; 3) $0.10 if the RT was within 100 ms of the standard. To ensure that participants also attended to the stop signals, the reward was only earned if they succeeded in stopping on 25% to 75% of the stop signal trials, in which case, they also earn $0.25 for meeting the stopping criterion.

Stopping cost and sampling-bias model simulation.

Our focus in Experiment 2 is on the selective stopping cost, the increase in RT that is observed for the invariant response on Selective Stop trials compared to Go trials (e.g., the foot response in the current study). Cost values greater than zero have been taken as indicative of a selective stopping cost in previous studies of selective stopping (Coxon et al., 2007; Majid et al., 2011; Aron & Verbruggen, 2008). We refer to this as the “restart model”, given the hypothesis that a generic stop command is initially applied to all prepared responses, followed by the restart of the invariant response (Figure 2.2C). The selective stop cost provides an estimate of the time required to restart the (transiently) disrupted invariant response. Note that the restart model is

inconsistent with the horse-race model, at least on successful stop trials. That is, restarting a planned response violates the independence assumption of the horse-race model.

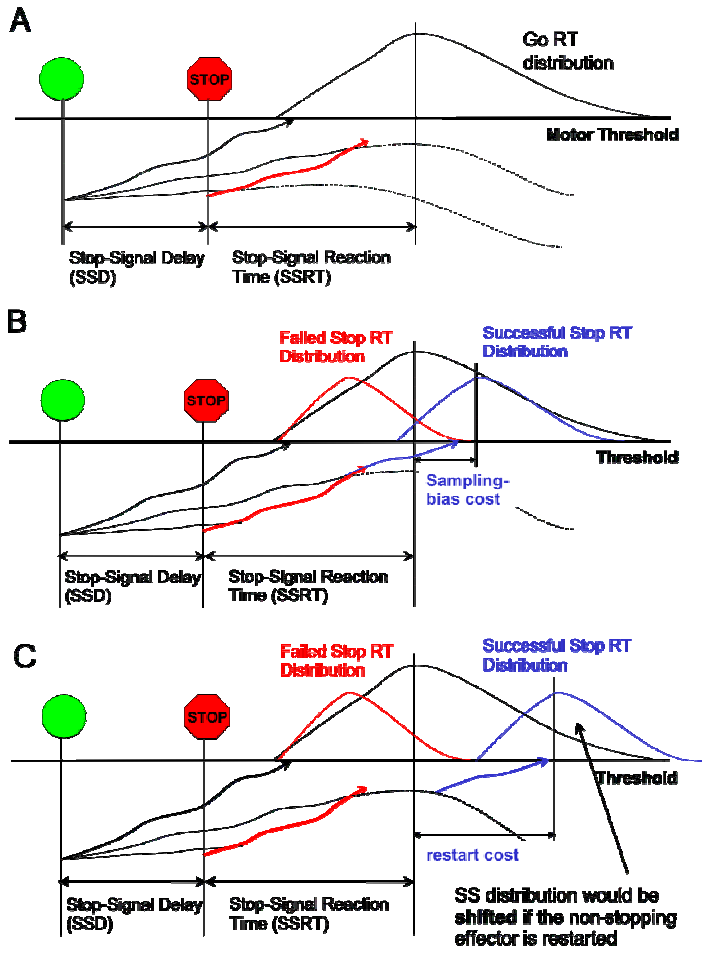

Figure 2.2. Horse-race model and the sampling-bias hypothesis for the residual cost. A. Graphic illustration of the horse-race model for stop-signal paradigm. B. Graphic illustration of the sampling-bias hypothesis. C. Graphic illustration of the re-programming hypothesis for stopping cost.

However a comparison of RTs on Selective Stop and Go trials is biased (Figure 2.2A). In the standard stop signal task (where a single response is prepared), RTs on trials in which the participant fails to stop are assumed to reflect the faster half of the RT distribution (Logan & Cowan, 1984). Indeed, it is this assumption which allows an estimate of the SSRT. In the selective stop task, we observe an overt response on all the stop trials (i.e., the foot response). By the horse-race model, the foot RTs from failed stop trials should provide an estimate of the faster half of the overall RT distribution. Moreover, foot RTs from successful stop trials should

be samples from the slower half of the distribution. Thus, a measure of selective stopping cost based on a comparison of foot RTs between successful Stop trials and Go trials would be biased since the former is drawn from only the slower half of the full distribution. We call this cost a *sampling-bias cost* (Figure 2.2B).

An unbiased estimate of the cost in selective stopping requires testing whether there is an additional increase in RT on successful stop trials, above and beyond the sampling-bias cost. Thus, our analysis focuses on asking whether the behavioral cost is significantly greater than modeled sampling-bias cost, an unbiased test of the restart model.

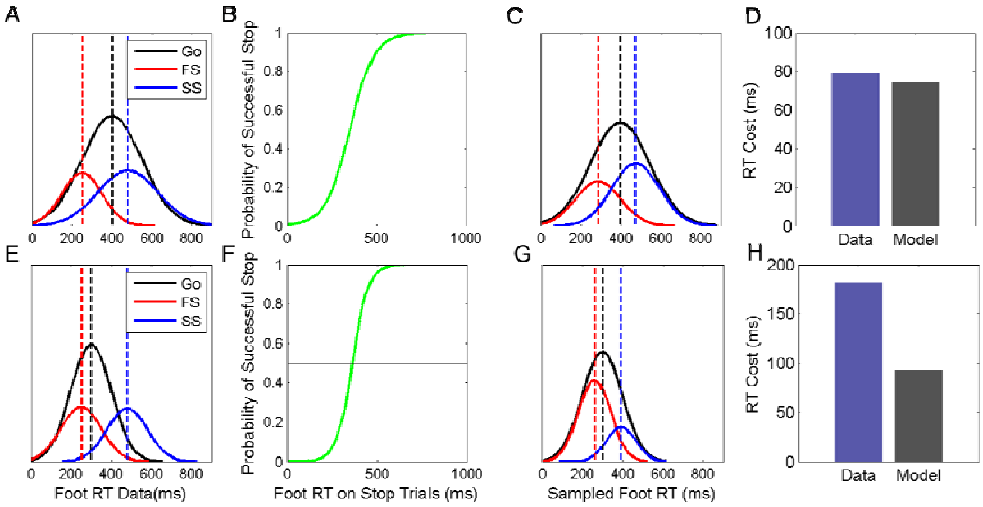

Figure 2.3. Sampling Bias Model's simulation to illustrate the case when sampling-bias cost can fully account for the behavioral cost (A-D), and the case when a restart-cost exist in the data so that the sampling-bias cost can not fully account for the behavioral cost (E-H). A, E. Simulated Go, Failed Stop (FS), and Successful Stop (SS) Foot RT data using normal distributions. B, F. stopping function *P*(*stop* | *foot RT*) using logistic regression to fit the FS and SS data sets shown in (A & E). C, G. Normal fits of a new set of samples drawn from the Go Distribution in (A & E), and FS and SS distributions classified using the stopping function in (C & F). D, H. Cost in the simulated data (A & E) compared to the modeled Sampling-Bias cost.

We used individualized stop probability functions in our Sampling Bias Model to estimate the sampling-bias cost (Figure 2.3B&F). We computed the stopping function *P*(*stop* | *foot RT*) using logistic regression on the foot RT data from the selective stop trials. This was done on the individual data sets, with separate functions computed for each of the four test sessions (300 observations/data set). We then fit the data from the Go trials with a Gamma distribution (Ratcliff & Murdock, 1976) and took 1000 samples from this distribution to define a foot RT sample (*foot RT_GO*). Using the function *P*(*stop* | *foot RT*), the samples were classified as successful stop (*foot RT_SS*) or failed stop (*foot RT_FS*) (Figure 2.3C&G). From these samples, we could estimate the sampling-bias cost as the difference between the simulated means of the

successful stop and go RTs [median(*foot RT_SS*) - median(*foot RT_GO*)]. Figure 2.3 illustrates two simulations in the case where the modeled cost can fully account for the behavioral cost (A-D), and where there is restart-cost in the data, so that the sampling-bias cost can not fully account for the observed cost (E-H).

2. 3 Results

2.3.1 Experiment 1

Experiment 1a.

Behavioral data.

The means for Go RTs, Failed Stop RTs, the estimated SSRTs, and the percentage of successful stops are presented in Table 2.1. These results are comparable to those reported in previous stop-signal studies (e.g., Logan et al., 1984). The tracking procedure successfully produced stopping rates near 50%. We performed three-way repeated-measures ANOVAs on the RT data with RT-type (Go vs. Failed Stop), session (practice vs. TMS), and finger (index vs. pinkie) as the three factors. As typically observed in stop-signal tasks, RTs on Failed Stop trials were faster than those observed on Go trials, ($F(1,13) = 57.68$, $p < 0.001$). A main effect of finger indicated that RTs were faster with the pinky compared to the index finger ($F(1,13) = 20.59$, $p < 0.01$). Importantly, there were no reliable differences between the practice and TMS blocks ($F(1,13) = 0.11$, $p = 0.74$), indicating that the TMS pulses did not produce a reliable change in RT, nor did this factor interact with RT-type ($F(1,13) = 2.12$, $p = 0.17$) or finger ($F(1,13) = 0.06$, $p = 0.82$). We also did a two-way repeated-measures ANOVA (TMS side × finger) on the Go RTs. RTs were not affected by whether or not the stimulation was applied over the contralateral or ipsilateral hemisphere to the response hand ($F(1,13) = 0.00$, $p = 0.997$).

However, applying TMS contralateral to the response hand did seem to affect SSRT ($F(1,13 = 4.71$, $p < 0.05$). SSRTs were delayed in index finger for about 20 ms, and in pinkie for about 16 ms.

CS excitability.

The data for one participant was excluded from the analysis because there were too few trials in which the TMS pulses fell within the criterion window of 120 – 20 ms before EMG onset.

The mean baseline MEP amplitudes for right FDI and ADM were 1.16 mV (SE = 0.21) and 0.33 mV (SE = 0.07), respectively (Table 2.2). Comparable values for left FDI and ADM were 0.94 mV (SE = 0.16) and 0.47 mV (SE = 0.15), respectively. The larger FDI values likely reflect the fact that we positioned the coil to maximize the MEP in FDI.

Our main interest is in the effect of response inhibition on task-relevant and task-irrelevant muscles. Table 2.2 presents the MEP data on Go and Successful Stop trials for selected and non-selected muscles in the responding hand, as well as homologous and non-homologous muscles in the resting hand.

	Left M1 Stimulation				Right M1 Stimulation					
	R Index (exp 1a)		R Pinkie (exp 1a)		L Index (exp 1a)		L Pinkie (exp 1a)		R Foot (exp 1b)	
	Practice	TMS	Practice	TMS	Practice	TMS	Practice	TMS	Practice	TMS
Go RT (ms)	414 (16.66)	424 (17.26)	388 (18.03)	398 (19.37)	426 (21.42)	423 (21.31)	407 (22.83)	399 (22.78)	571 (16.41)	579 (20.74)
Failed Stop RT (ms)	359 (11.34)	346 (14.31)	333 (12.24)	328 (11.74)	376 (18.42)	378 (17.92)	360 (14.82)	353 (16.36)	510 (14.94)	495 (19.10)
SSD (ms)	171 (13.51)	169 (18.94)	155 (10.73)	155 (14.79)	184 (16.85)	188 (22.30)	177 (17.32)	172 (18.63)	281 (13.31)	282 (18.75)
SSRT (ms)	243 (8.05)	255 (11.63)	232 (10.15)	243 (9.54)	242 (9.70)	235 (10.65)	230 (9.04)	227 (9.35)	290 (9.33)	297 (13.13)
Percentage successful stop	0.51 (0.02)	0.56 (0.04)	0.49 (0.02)	0.50 (0.03)	0.50 (0.02)	0.52 (0.04)	0.50 (0.02)	0.46 (0.03)	0.53 (0.01)	0.60 (0.03)

Table 2.1. Behavioral results for experiment 1a and 1b. Mean (SE) RT for go and failed stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop.

		Right (Task-relevant Hand)		Left (Resting Hand)	
		FDI	ADM	FDI	ADM
Baseline (n = 14)		1.14 (0.20)	0.34 (0.07)	0.92 (0.15)	0.45 (0.14)
Selected / Homologous	Go (n = 13)	1.75 (0.20)	0.61 (0.09)	0.79 (0.14)	0.47 (0.16)
	Failed Stop (n = 6)	2.50 (0.35)	0.81 (0.17)	0.91 (0.19)	0.56 (0.19)
	Successful Stop (n = 13)	1.03 (0.13)	0.34 (0.07)	0.64 (0.12)	0.40 (0.16)
Non-selected / non-homologous	Go (n = 14)	0.50 (0.11)	0.18 (0.03)	0.49 (0.08)	0.40 (0.17)
	Failed Stop (n = 10)	0.39 (0.09)	0.22 (0.04)	0.49 (0.08)	0.36 (0.14)
	Successful Stop (n = 14)	0.49 (0.11)	0.22 (0.04)	0.52 (0.08)	0.44 (0.20)

Table 2.2. Mean (SE) amplitude (in millivolts) of MEPs recorded from each muscle in experiment 1a (N = 14).

We first ask how MEPs change on Go trials compared to baseline. There was a significant increase in MEPs in the selected effector ($t(12) = 5.42$, $p < 0.001$). The opposite pattern is observed in the non-selected effector: Here the MEPs showed a significant reduction from baseline ($t(12) = -7.30$, $p < 0.001$). This pattern is consistent with preparatory processes that both facilitate the selected effector and inhibit the non-selected, but task-relevant effector (Sohn, et al., 2003; Duque et al., 2010). Interestingly, we also observed a significant reduction from baseline in the muscle homologous to the non-selected effector ($t(12) = -3.80$, $p < 0.005$). For example, when the right pinkie was cued to move, inhibition was evident in both the right and left FDI, even though the left hand was task irrelevant. A similar effect was not observed in the resting finger homologous to the selected response ($t(12) = 0.12$, $p = 0.91$).

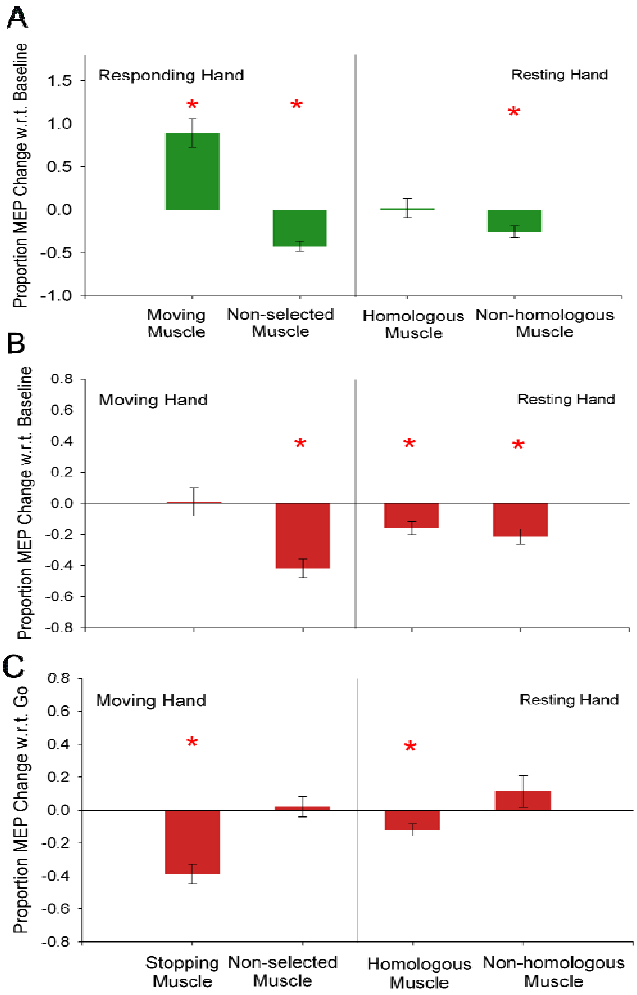

Figure 2.4. Percentage change in the amplitude of MEPs in four functional classes of muscles in Experiment 1a. A. On a go trial compared with baseline; B. On a successful-stop trial compared with baseline; C. On a successful-stop trial compared with go.

We next examine the changes in the MEPs following the onset of a stop signal[1]. Relative to the Go trials, there was a dramatic reduction in the MEPs in the selected muscles on successful stop trials ($t(12) = -6.54$, $p < 0.001$). It is possible that this reduction might reflect an absence of preparatory processes in the cued effector, as there was no difference in MEP values between Successful Stop trials and Baseline ($t(12) = 0.10$, $p = 0.92$). However, this value more likely reflects the operation of active inhibition triggered by the stop signal, as we did not observe further reduction of MEPs in the non-selected and the non-homologous muscles.

The effects of the stop signal on CS excitability in the non-selected effector and task irrelevant, resting hand bear on the question of the selectivity of inhibitory commands. There were no differences in the MEPs for the non-selected effector on Successful Stop trials compared to the Go trials ($t(12) = 0.34$, $p = 0.74$). Thus, while this effector was inhibited as part of the response selection process (e.g., relative to baseline), there is no indication of further modulation of CS excitability in this finger by the stop signal.

The picture is more complicated when examining the data for the task irrelevant, resting hand. For the non-homologous effector (e.g., left pinky when right index finger was cued), there is no change in the MEPs ($t(12) = 1.25$, $p = 0.23$). However, for the homologous effector (e.g., left index when right index was cued), on a successful stop trial, there is a reliable drop in the MEPs, an effect that is reliable when compared to Go trials ($t(12) = -2.98$; $p < 0.05$) and also falls below the values observed at baseline ($t(12) = -3.74$, $p < 0.005$).

In summary, the results are inconsistent with the hypothesis that processes recruited to abort a planned response have a generic inhibitory effect on the motor system. This conclusion is most clearly supported in the data for the non-selected effector. MEPs for this effector were strongly attenuated following the initial Go stimulus as part of the response selection process (e.g., relative to baseline). However, there was no indication of further modulation of CS excitability in this finger by the stop signal. Similarly, we did not observe any indication of stop-signal related inhibition in the non-homologous muscle of the resting hand. However, we did observe some spill-over of inhibition in the resting hand muscle homologous to the cued effector. Inhibitory signals, either related to response selection (Go trials) or stop signals were evident in homologous muscles of the resting hand. We believe that this signal is more related to the special mechanical properties of the homologous muscles than an evidence of generic inhibition (see Discussion).

Experiment 1b.

In Experiment 1b participants performed a two-choice stop signal task, responding on one of two response keys with the foot. The behavioral results were comparable to Experiment 1a (Table 2.1). Failed Stop RTs were faster than Go RTs, ($F(1,15) = 107.03$, $p < 0.001$) and TMS did not affect RT ($F(1,15) = 0.11$, $p = 0.75$) nor SSRT ($F(1,15) = 0.37$, $p = 0.55$).

TMS was applied over right motor cortex to probe corticospinal excitability in the left hand in the task. MEP values did not differ between the first and second baseline measurements, with mean values across the two baselines of 0.76 (SE = 0.08) for FDI and 0.28 (SE = 0.03) for ADM.

Given that there was no difference between the two baseline probes, ($F(1,15) = 0.04$, $p = 0.85$), we combined the data across the two baselines in the analysis of the MEP data (Figure 2.5).

The MEPs in both left FDI and ADM did not show any significant change from baseline on Go trials when the foot pressed the key (both p's > 0.1). More important, no change was observed in either hand muscle when a prepared foot response was aborted on successful stop trials (FDI: $t(15) = -1.63$, $p = 0.12$; ADM: $t(15) = 0.47$, $p = 0.64$).

In summary, the results from both Experiment 1a and 1b fail to support the generic inhibition hypothesis.

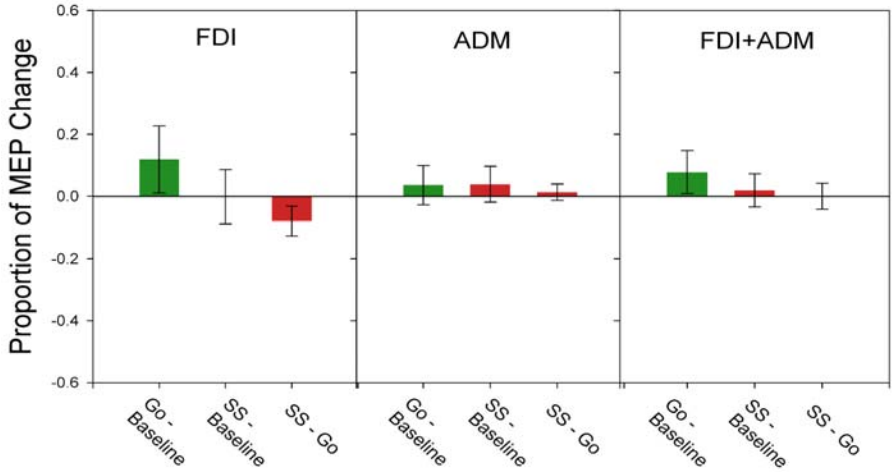

Figure 2.5. Proportion change in the amplitude of MEPs in left FDI, ADM, and FDI and ADM combined in Experiment 1b, on Go trials, Successful Stop trials relative to the two baselines combined, and on Successful Stop trials relative to Go trials.

2.3.2 Experiment 2

In Experiment 2, Trials in which the RTs for the finger and foot responses were more than 100 ms apart were considered as desynchronization errors. Some of these occurred during the first few blocks of training, but were rare in latter blocks (approximately 1% in both visual and tactile conditions). Participants all learned to perform the primary choice task quickly and maintained a high level of accuracy across the entire experiment, with Go trial accuracy at 96% (SE = 0.003), and on average, each person earned about $3.49 bonus per session.

A summary of the RT data and measures of stop signal performance is presented in Table 2.3. Foot RTs were slower than manual RTs ($F(1,9) = 17.31$, $p < 0.005$), but the mean difference was 28 ms (SE = 6.6), indicating that the two responses were initiated in a synchronized manner.

Means RTs were relatively fast and improved over the two days of training. For the visual condition, manual RTs averaged 320 ms (SE = 15) on Day 1 and 296 ms (SE = 12) on Day 2.

For the tactile condition, the means were 307 ms (SE = 17) and 280 ms (SE = 10) for Days 1 and 2, respectively. The reduction across days was reliable ($F(1,9) = 10.30$, $p < 0.05$) in four-way repeated-measures ANOVAs (Effector × RT Type × Signal Type × Day). RTs were also faster on Failed Stop trials compared to Go trials ($F(1,9) = 32.15$, $p < 0.001$). The effect of Signal Type was not reliable ($F(1,9) = 1.02$, $p = 0.34$).

	Color		Tactile	
	Day 1	Day 2	Day 1	Day 2
Go RT hand (ms)	319 (15.23)	296 (11.51)	309 (21.59)	276 (11.91)
foot (ms)	344 (19.62)	315 (15.04)	332 (20.12)	302 (12.64)
Failed Stop RT hand (ms)	290 (10.64)	272 (10.41)	279 (13.79)	256 (8.54)
foot (ms)	327 (14.03)	303 (13.47)	309 (12.83)	288 (10.46)
SSD (ms)	106 (13.30)	88 (6.49)	106 (11.42)	92 (6.32)
SSRT (ms)	213 (8.77)	208 (9.70)	204 (12.70)	184 (8.79)
Percent successful stop	0.38 (0.03)	0.38 (0.02)	0.39 (0.02)	0.38 (0.02)

Table 2.3. Behavioral results for Experiment 2 (N = 10). Mean (SE) hand and foot RT for go and failed stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop.

In general, performance on the selective stop task was similar to that observed in the standard "all" stop task (Experiment 1a). The SSD averaged 98 ms (SE = 8) and did not differ between the visual and tactile conditions ($F(1,9) = 0.04$, $p = 0.84$), nor vary across days ($F(1,9) = 3.88$, $p = 0.08$). The SSRT was faster to the tactile stop signal and showed a reduction across days, but neither effect was reliable (Signal Type: ($F(1,9) = 2.35$, $p = 0.16$; Day: $F(1,9) = 4.76$, $p = 0.06$). The percentage of successful stops averaged only 38% in this experiment, a value lower than that observed in Experiment 1. We assume this reflects that it is more difficult to selectively stop one movement within a compound response, especially given our monetary incentive system.

Foot RTs were slower on successful Stop trials compared to Go trials ($F(1,9) = 28.61$, $p < 0.001$) (Figure 2.5), the signature of a selective stopping cost (Coxon et al., 2007; Majid et al., 2011; Aron & Verbruggen, 2008). To evaluate how this cost was influenced by the manner in which the stop signal was cued as well as by practice, we performed a two-way repeated-measures ANOVA with the factors Signal Type (Visual vs. Tactile) and Day (1 vs. 2). The cost was larger for the visual stop signal (67 ± 15 ms) compared to the tactile stop signal (42 ±10 ms), but the effect was not reliable ($F(1,9) = 2.73$, $p = 0.13$). The selective stop cost was reduced with practice ($F(1,9) = 12.72$, $p < 0.01$), dropping from 67 ms (SE = 12) on Day 1 to 41 ms (SE = 9) on Day 2. The interaction of these two factors was not reliable ($F(1,9) = 0.14$, $p = 0.71$).

Importantly, the size of the cost is significantly greater than zero for both signal types and during the second session (Day 2, collapsed over signal types: $t(19) = 4.438$, $p < 0.001$). Based on this analysis, one would conclude that selectively stopping one movement within a compound response imposes a cost on the invariant response, a central assumption of the reprogramming hypothesis. However, as noted above, a comparison between foot RTs on successful Stop and Go trials is biased since the former is composed of the slower half of the full RT distribution. Given this, we used the stop probability functions and Go RT distributions for each participant to estimate the size of this bias (see Methods). We compared this estimate to the observed cost.

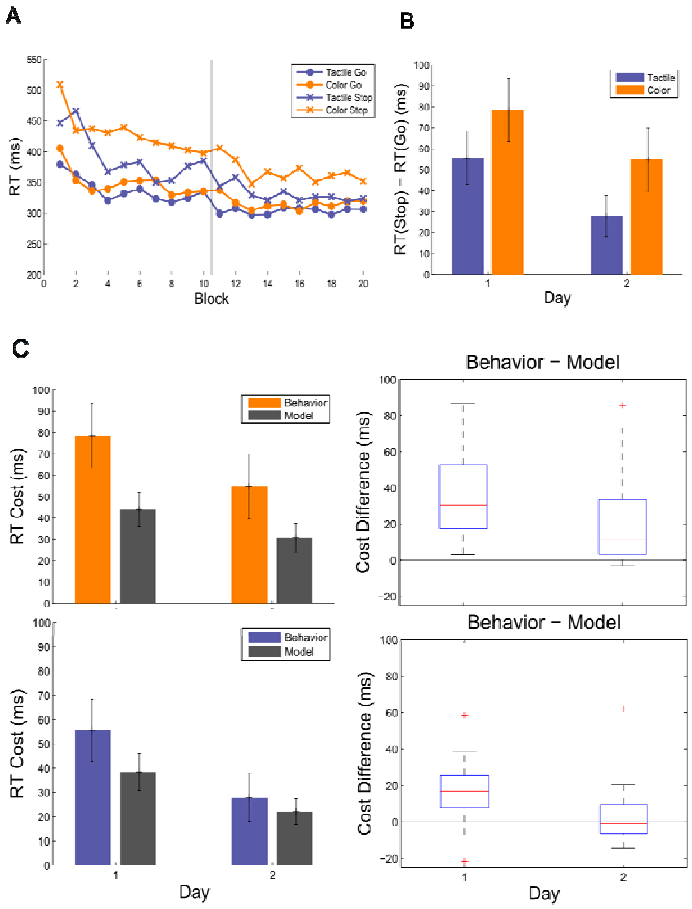

Figure 2.6. Non-stopping effector, foot RT and selective-stopping cost data for experiment 2. A. Foot RTs by block. B. Cost by day. C. Estimated residual cost with Sampling-bias Model 1, compared to behavior cost.

Figure 2.5C shows the mean cost predicted by the Sampling Bias Model for each stop signal and session. For the visual condition, the observed and modeled costs was significantly different

on Day 1 ($t(9) = 3.81$, $p < 0.005$) and marginally significant on Day 2 ($t(9) = 2.18$, $p = 0.06$). For the tactile condition, the difference was significant on Day 1 ($t(9) = 2.49$, $p < 0.05$). However, on Day 2, the cost difference was not reliably different than zero ($t(9) = 0.65$, $p = 0.53$), indicating that the participants were successfully able to abort one component of a planned action without a corresponding cost in a second component of that action.

In summary, a simple analysis based directly on the RT data suggests that people are unable to selectively stop one response within a multi-component action. However, this analysis is based on a biased sampling procedure. When this bias is accounted for, the results indicate that, the use of a salient stop signal (tactile), coupled with a modest degree of practice, led to the complete elimination to a cost associated with selective stopping. Participants were able to produce the foot response independent of whether they were simultaneously initiating or aborting a planned manual response.

2. 4. Discussion

In the present study, we investigated people's ability to selectively stop. Previous studies have found that stopping one component of a compound action interferes with the not-to-be-stopped components, resulting in a RT "stopping" cost (Coxon et al., 2007; Majid et al., 2011; Aron & Verbruggen, 2008). These findings have lead to the hypothesis that the hyper-direct pathway, mediated by the STN, is always involved in selective stopping. The inference is that, because of the poor resolution of this system, people have to globally inhibit the entire action and then restart the non-stopping component. With two experiments, we provide converging evidence that people can selectively stop. In Experiment 1, we used single-pulse TMS to probe participants' CS excitability when they performed a stop-signal task. In Experiment 1a, the muscles controlling the non-selected response in a task-relevant hand, as well as non-homologous muscle in a task-irrelevant hand did not show inhibition associated with the stop-signal. In Experiment 1b, when people were performing the stop-signal task with their foot, we also did not observe inhibition related to the stop-signal. In Experiment 2, we trained participants to do a selective stop task with a salient stop signal (tactile). Participant were able to reduce the selective stopping cost in the non-stopping effector, and our Sampling Bias Model showed that the residual cost was a statistical artifact of the stop-signal design.

2. 4.1 Psychological issues on selective stopping

Parameters that affect the Sampling Bias Model.

The Sampling Bias Model's prediction of the cost is sensitive to two sets of parameters: the means of Go RT and Successful Stop RT distributions, and the variances of these two distributions. When the mean of Successful Stop RT distribution is shifted to the right, outside

the range of the Go RT distribution, the observed stopping cost will increase. This is the case when the horse-race model is violated, such as when a restart process is required (Figure 2.3E-H). The modeled sampling-cost, however, will remain to be small, as the Go Distribution is not altered.

The variance in Go RT distribution does affect the predicted cost since the modeled cost will also increase. When the full distribution is spread out, the difference between the mean values of the Go RT and Successful Stop RT distributions grows larger. However, this increased cost need not be due to a restart process. It may simply be an artifact of the sampling bias in the task.

One advantage of the model is that by using individualized stopping function, we are able to bypass the assumption of a 50% stopping rate. These assumptions are used in the horse-race model to calculate SSRT (Logan and Cowan, 1984). However, this rate may not be empirically obtained, and therefore may affect the model's prediction of the sampling-bias cost. However, one caveat of using individualized stopping functions is that when the independent assumption of the horse-race model is violated and there was is a restart cost, the inhibition function will be "contaminated," shifting the stopping function curves slightly to the right. This will end up with more trials being labeled Failed Stops than Successful Stop in the new samples from the Go RT distribution (Figure 2.3G). This will lead to a slight overestimate of the sampling-bias cost because the mean of the new set of Successful Stop RTs will be shifted to the right. However, this issue is not detrimental to the model's ability to discriminate between the restart hypothesis and the sampling-bias hypothesis. The reason is that the new set of data are solely sampled from Go Foot RT distribution. Therefore the amount of overestimation will be minimal and always within the range of the original full distribution. In contrast, the restart hypothesis dictates that the Successful Stop RT distribution will be shifted outside the Go RT distributions (Figure 2.3E). In sum, we are safe to claim that when this happens, the model is still accurate in providing an unbiased test of selective stopping cost.

Saliency of Stop-signal and the Assignment Problem in Selective Stop.

In Experiment 2 we showed that people benefited on a selective stop task when using a salient stop signal (tactile) and were given extended training. In the psychological literature, it has been debated whether one can learn to control a stop signal. For example, Logan et al. (1984) showed that the SSRT and inhibition functions were invariant to the type of primary task (e.g., simple RT or choice tasks). The question is then what was learned in the Selective Stop task?

Schneider and Shiffrin (1977) proposed the idea of *consistent* vs. *varied mapping* in the development of automaticity. Automaticity with practice is only possible with consistent mapping, in which the same stimulus and response are constantly associated with each other. In varied mapping, the stimulus is mapped onto different responses and therefore practice does not improve performance. The stop signal would be a case of varied mapping, since the same choice stimulus is sometimes associated with go and sometimes associated with stopping. Verbruggen & Logan (2008) found that SSRT can be improved with training, but only with continuous stimuli-stopping mapping, not varied mapping. However, our results show that training can reduce the Selective Stop cost. Not only did RTs improve for both the stopping hand response

and the invariant foot response, but the SSRT also improved and the stopping cost became smaller.

This finding poses a challenge to the consistent vs. varied-mapping notion. In the Selective Stop task, not only do Go and Stop trials share the same choice stimuli, but the secondary task, stopping, also has two conflicting components that share the same stimuli (the stop signal). This is a further layer of varied mapping. Thus, training may not only automatize the invariant response, but may also help automatize the response to the stop signal. The key of this improvement may be in the separation of the two components with training. Our tactile condition strongly argues in favor of this notion. Moreover, we provided incentives to the participants to maintain their speed on the invariant response. Both of these experimental manipulations argue for a separation of mappings rather than for a strengthening of associations.

The common assignment problem in different inhibition situations.

Combining the stop-signal task with a flanker task or Simon task, Verbruggen et al. (2005) (also Kramer et al., 1994; Riderinkhof et al., 1999) showed that stimulus-stimulus and stimulus-response compatibility can affect how well people stop. In the flanker and Simon tasks, an RT cost is observed when stimulus-stimulus or stimulus-response mappings are incongruent in the flanker task (Eriksen and Eriksen, 1974) and Simon task (Simon & Rudell, 1967). Kramer et al. (1994) (also Ridderinkhof et al., 1999; Verbruggen et al., 2004) showed that SSRT increased when responding to a stop signal presented in an incompatible flanker condition. Verbruggen et al (2005) found similar effects when combining a stop signal task with a Simon task. A generally acknowledged notion in this line of research is that response suppression in the primary task (Go) interferes with the inhibition process required for stopping. Ridderinkhof (2002) suggested that this was because that the two kinds of inhibition share common resources/pathways.

The overlap of pathways and competition for resources at different stages of processing is a common theme in research on cognitive control (Broadbent, 1958; Kahneman, 1973; Keele and Neill, 1978; Wikens, 1984). Competition in dual tasks is often reflected in a psychological refractory period (Kahneman, 1973; Kantowitz, 1974). The stop-signal task can also be viewed as a dual task in which the second response is to inhibit the first response. In the Selective Stop task, the stopping process can be viewed as another form of a dual task: Going with one effector while simultaneously stopping another one. We propose that the selective stopping cost may be accounted in the similar way as an assignment problem: resolving a conflict either in stimulus-stimulus or stimulus-response mappings. The amount of overlap in the conflicting stimuli or response can account for the stopping cost. The more overlap, the more conflict to resolve. The difference, though, is in the stage of overlapping. In flanker and Simon tasks, the assignment problem is at the earlier stage (Go process), whereas in the selective stopping situation, the assignment problem is at the later stage (Stop process).

One challenge to this hypothesis comes from the work of Logan and Burkell (1986) who found relatively smaller interference in a stop-signal task compared to a dual-task. We think this difference could still be accounted for by the amount of overlap. In the stop-signal task, the go

component entails an overt response, whereas the stopping component does not. In contrast, in standard dual-task experiments, there are always two overt responses. Using a sequential stop-signal design, Yamaguchi et al. (2011) examined the psychological refractory period (PRP) effect of on SSRTs in two stop-signals following one another, and found no evidence of interference on SSRT. However, the cost on the second response RT was still quite large.

2.4.2 When is global and when is local?

We showed that people can selectively stop. This raises the question of when are generic or selective stop signals recruited, and do the two kinds of inhibition mechanisms interact? DeJong et al. (1995) proposed that a subcortical, generic mechanism operates quickly to prevent output that might occur if one is dependent on a slower, cortical-based selective mechanism. On the other hand, a central, selective mechanism may be critical in our daily life since it allows fluent actions. This notion goes along well with the idea of bottom-up vs. top-down inhibition (Ach, 1935; Bargh & Ferguson, 2000; Miller, Galanter & Pribram, 1960). Based on his ERP findings, DeJong suggested that the peripheral mechanism may work downstream, while the central mechanism operates upstream.

Strategy and training might be used to manipulate the two kinds of mechanisms. The results of Experiment 2 show that people can be trained to selectively stop. We postulate that training may promote the fine-tuned separation of motor mappings. Another possibility as DeJong et al. (1995) suggested is that the central mechanism activates a large motor map to facilitate concurrent or alternative movements.

However, this training may be constrained physically. In Experiment 1A, we observed a spill-over of inhibition in muscles homologous to the selected muscle. More interestingly, this spill-over was observed in two ways. First, there was inhibition in the non-selected muscle when preparing a response for the Go signal. Second, there was inhibition in the selected muscle following the stop signal. We believe that this the inhibition of the former type is not evidence of a generic stopping command, but rather is related to the mechanical wiring of homologous muscles (Swinnen et al., 1997; Serrien & Swinnen, 1997). The locus of this a constraint involving cross-talk between homologous muscles may arise from subcortical mechanisms.

The spillover of inhibition during Go trials on both the non-selected muscle and non-homologous muscle replicates Sohn et al.'s (2003) finding. However, in contrast to that study, facilitation was not present in homologous muscles in our experiment. This discrepancy could be due to the sensitivity of our single-pulse TMS method, but more interestingly, the timing of the pulses could be relevant. Our window of stimulation was before EMG onset, whereas Sohn et al. looked at facilitation after EMG onset. Thus, the inhibition we observe can only be attributed to movement preparation, while the inhibition in Sohn et al. may be related to processes recruited for movement initiation and execution.

2.4.3 Neural mechanisms of the global and local inhibition

Cortico-basal-ganglia-thalamus pathways have been suggested as playing a central role in action selection (Alexander et al., 1990; Mink, 1996). Several of these pathways have also been implicated in inhibitory control, specifically, the hyper-direct and indirect pathways of the basal ganglia.

In addition to these subcortical mechanisms, several studies also emphasize a contribution of right inferior frontal gyrus (IFG) and pre_SMA in inhibition (e.g. Aron & Poldrack, 2006; Aron et al., 2007; Derrfuss et al, 2004). Moreover, STN activation during stopping is not always observed (Xue et al., 2008). The inconsistency of these findings led us to suspect that different pathways may be differentially involved in global and local inhibition, and to question whether there is a mechanism that produces generic stop commands.

As shown in our results, training can reduce the Selective Stop cost. There are two apposing view of how learning migrates between BG and PFC. One view emphasizes that the basal ganglia are critical for habit formation (see a review by Yin and Knowlton, 2006). A second emphasizes that over the course of learning, control shifts from the basal ganglia to the cortex (reviewed in Ashby et al., 2010). In the former view, the learning to selectively stop may involve the initial activation of generic subcortical pathways (such as the hyperdirect pathway of the basal ganglia), but with training, shift a more specific pathway (such as the indirect pathway of the basal ganglia). The latter hypothesis would suggest that selective stopping after training may be achieved by pure cortical-based inhibition, bypassing the basal ganglia.

One way to resolve the two conflicting views might be drawn from the literagture concerning the role of anterior cingulate cortex (ACC) in conflict monitoring (Botvinick et al., 1999, 2001). One model of ACC function in conflict monitoring builds on the idea that ACC is more active with higher level of conflict, as well as after the occurrence of error. If this also occurs during selective stopping, activation of the ACC may be correlated with the reduction of Selective Stop cost.

2.4.4 Conclusion

In two experiments we showed that people can selectively inhibit some component of an on-going action. In Experiment 1, we used TMS to probe the corticospinal excitabilities during a stop-signal task, we did not find extra inhibition in the selected muscles in the task-relevant hand, non in the task-irrelevant hand. In Experiment 2, we show that with a moderate amount of training, stop cues salient to the stopping effect, and incentive to speeding up non-stopping effector, people can learn to selectively stop. Simulations using the Sampling Bias Model also show that the residual cost in the non-stopping effector can be explained by a statistical artifact inherent in the standard stop-signal task. Our findings as a package strongly argues against the hard-wired generic stopping mechanism that can not bypass, but supports the idea the constraint in selective stopping is soft and can be overcome with training.

Footnote

1. We note that the MEPs were larger for the selected effector on these failed stop trials compared to Go trials (Table 2.2), consistent with the hypothesis that elevated motor cortex excitability reflected more advanced motor preparation (and thus, the failure to abort the response). However, because the majority of Failed Stop trials were faster than Go trials, many of the TMS pulses occurred after EMG onset and we had fewer than eight observations for eight the participants. For this reason, we did not statistically analyze the Failed Stop data.

Chapter 3

Intrinsic Fluctuation in Inhibitory Control

3.1 Introduction

Psychological and neuroscientific studies generally emphasize the brain as an information processing system, taking input from the external world and emitting output after certain forms of information processing (Marr, 1982). As such, research often focuses on stimulus-driven activity and responses: We manipulate a stimulus and observe the response to make inferences about internal processing. However, it is widely recognized that many aspects of mental activity are less correlated with external stimulation (James 1890).

Indeed, there has recently been a flurry of interest in these internal processes. For example, fMRI research on the "resting state", traditionally used as a baseline for comparison with experimentally-induced states, has become a topic of considerable interest effects (reviewed in Fox & Raichle, 2007). More generally, correlations among the intrinsic fluctuations in different neural networks have been hypothesized to reflect the functional topography of the brain (reviewed in Bullmore & Sporns, 2009). Variation in resting state activity has been associated with a wide range of behavioral effects with the basic assumption that this activity reflects both trait and state factors. For example, from the resting-state activities, Seeley et al. (2007) identified a "salience network" that correlated with individual pre-scan anxiety scores, whereas an "executive-control network" that correlated with individual performance on an executive-control task, Trail-Making Task. Also, fluctuations in the activity of sensorimotor cortex correlate with inter-trial variability observed on a force production task (Fox et al., 2007).

In many of these studies, measures of intrinsic activity are obtained at a time distant from the behavior/trait measure that is subsequently correlated with this activity. One can also ask how moment to moment fluctuations in internal states influence behavior on a more microscopic time scale. For example, in an event-related fMRI study, Hesselmann et al. (2008) showed that the fluctuations in the motion-sensitive area hMT+ correlate with trial-by-trial motion detection behavior. In the present study, we adopt this approach to explore how local fluctuations in cortico-motor pathways influence people's ability in inhibitory control.

The *stop-signal* task, in which a delayed stop signal provides a cue that a planned response should be aborted, has been extensively used to study inhibitory control. The data from this task

have been taken to indicate the operation of a “horse race”, entailing a race between processes devoted to generating the cued response and processes devoted to ensuring that this response is aborted (Lappin & Erikson, 1966; Logan & Cowan, 1984; Logan et al., 1984; Logan, 1994; Verbruggen & Logan, 2008). This model is a classic example of an externally-driven process: on Stop trials, the Go and Stop signals trigger a set of internal processes, and it is the processing times of these stimulus-based events that will determine if the response is appropriately aborted. The functional role of intrinsic fluctuation on inhibitory control has not been explored.

The state of the motor system prior to the onset of the trial may be relevant when considering whether or not a person is able to abort a planned response. Specifically, we hypothesize that it may be more difficult to inhibit a planned action when excitability is high, whereas inhibition may be more easily achieved when excitability in low. The final behavioral outcome may be the result of the interaction of intrinsic and extrinsic processes. Moreover, the case of failed inhibition in the stop-signal task offers us an opportunity to examine the interplay of the extrinsic and intrinsic activities, because after encountering a failed inhibition, the two dynamics should work towards opposite directions. Specifically, while the intrinsic system was at a heightened state, the extrinsic control system might work to exert increased inhibition.

We used transcranial magnetic stimulation (TMS) to assess excitability of the corticospinal (CS) pathway during the stop-signal task. In traditional single-pulse TMS studies, the baseline level of CS excitability is assessed as motor-evoked potentials (MEPs) during the inter-trial-interval (ITI), and this baseline is used to compare with MEPs during the period of task performance (e.g., Duque & Ivry, 2009). In the current study, we focus on probes of CS excitability during the ITI instead. In three experiments, we examined if the level ITI MEPs would predict people's performance on a stop-signal task, using either unimanual task, bimanual task, or both. We further look at the interaction of control process and intrinsic fluctuation by focusing on the sequential analysis of the aftereffect of behavioral outcome on MEPs. The results show a two-way interaction of the extrinsic and intrinsic processes in inhibitory control with stronger influence of extrinsic control processes in the bimanual task.

3.2 Materials and Methods

3.2.1 Participants

A total of 36 right-handed healthy participants participated in Experiments 1, 2, and 3. Handedness was confirmed via a condensed version of the Edinburgh Handedness Inventory (Oldfield, 1971). All participants gave written informed consent and were financially compensated for their participation. The protocol was approved by the institutional review board of the University of California, Berkeley.

3.2.2 Experimental Task.

Changes in corticospinal excitability were monitored via TMS over motor cortex while participants performed a stop-signal task. In Experiment 1, participants (4 women, 5 men; 22.8 ± 7.8 years old) performed a two-choice RT task, activating a response key by either abducting the right index finger (FDI) or flexing the right pinky finger (ADM). Each trial began with the presentation of a fixation marker, an asterisk, at the center of the screen for 500 ms (Figure 3.1A). This was followed by the go signal, a green circle that appeared on either the left or right side of the computer screen (10 deg eccentricity). Participants were instructed to make a speeded response with the spatially compatible effector (e.g., index finger when stimulus was on the left side). On a fixed percentage of the trials (33% during training and 40% during the TMS sessions), the green color changed to red, the stop signal. The interval between the onset of the green circle and the color change on stop trials, stop-signal delay (SSD), was determined in an adaptive manner (see below). On these trials, participants were instructed to try and abort the planned response. The circle remained visible for 1000 ms after the SSD, same for go and stop signals, and the screen then went blank for 2000 ms (inter-trial interval). The task-irrelevant, left hand rested on a pillow during the entire experiment.

An adaptive staircase procedure was employed to determine the SSD on a trial-by-trial basis. The SSD value was increased by 50 ms after a successful stop and decreased by 50 ms after a failed stop. This method converges to a SSD value at which participants will successfully stop a planned response on approximately 50% of the trials (Levitt, 1971, Osman et al., 1986, 1990, Band et al., 2003). Separate staircases were run for index and pinky finger trials given that the reaction time for these effectors may not be equivalent. Moreover, for each finger, there were two separate staircases, one starting at 50 ms and the other at 300 ms. Participants were informed of the general nature of the adaptive procedure and that they should expect to fail to stop on a large number of trials. They were instructed that the procedure was designed to minimize the effectiveness of adopting a strategy of delaying the response to determine if the stop signal would appear.

Each participant completed three training blocks, each composed of 48 Go trials and 24 Stop trials. The SSD estimates at the end of each training block were used as the initial values for the subsequent training block. This was followed by eight test blocks, four of which involved TMS over the right motor cortex and four of which involved TMS over the left motor cortex. The order of stimulation side was counterbalanced across participants. The mean SSDs in the last training block, one for each finger, were used as fixed SSDs in the test blocks. We opted to keep these values fixed because changing the SSD might affect our assessment of the influence of intrinsic fluctuations in cortical excitability. To prevent participants from adjusting to this fixed SSD, eight "catch" trials were included, four in which the SSD was reduced by 80 ms and four in which the SSD was increased by 80 ms. Thus, there were a total of 80 trials per test block, 48 Go and 32 Stop. One TMS pulse was applied on each trial, occurring 30 ms before the onset of the fixation marker.

Experiment 2 (8 women, 7 men; 25.1 ± 5.9 years old) followed the same design as Experiment 1, except that the choice in the stop-signal task was between the left and right index

fingers (Figure 3.1B). This change was made to allow us assess the relationship between cortical excitability and performance when the measures were obtained from the same hand or from different hands. In Experiment 3 (6 women, 6 men; 21.3 ± 1.5 years old), the within- and between-hand manipulations were performed in the same individuals in a single session. This provided a replication of Experiments 1 and 2. To complete testing within a single session, TMS was only applied over left M1. We focused on the same effector, right index, in the two tasks. Participants performed 4 blocks of the within-hand task (as in Experiment 1) and 4 blocks of the between-hand task (as in Experiment 2). There were three training blocks before each task to determine the SSD values.

Performance on the stop-signal task was assessed with the reaction time (RT) data and stop-rate data. RTs were recorded by the activation of microswitches on the response key board. We also estimated the stop-signal reaction time (SSRT), subtracting the mean of the SSD that led to a 50% stopping rate from the median of the Go RT distribution (Logan & Cowan, 1984).

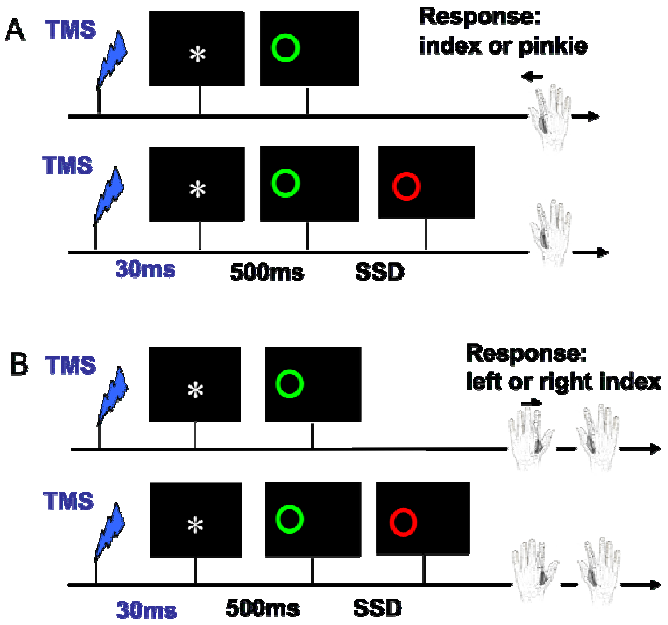

Figure 3.1. Stop-signal task procedure and TMS timeline used in A. Experiment 1 and 3a, within-hand task. B. Experiment 2 and 3b, between-hand task.

3.2.3 Assessment of cortical excitability.

Motor-evoked potentials (MEPs) were elicited by single-pulse TMS applied over the left or right primary motor cortex (M1). The TMS pulses were applied with a figure-of-eight coil (diameter of wings 70mm) connected to a Magstim 200 magnetic stimulator (Magstim, UK). Participants

wore a tight-fitting electroencephalography (EEG) cap and the coil was placed tangentially on the scalp. The handle was oriented toward the back of the head and laterally at a 45° angle from the midsaggital line, approximately perpendicular to the central sulcus. The position of the stimulator on the skull was adjusted to produce a maximal response from the first dorsal interosseous muscle (FDI). At this location, suboptimal responses are also observed in the abductor digiti minimi muscle (ADM). The location was marked on the EEG cap to provide a reference for the experimental session.

The resting motor threshold (rMT) was defined as the minimal TMS intensity required to evoke MEPs of ~50 μV (peak-to-peak amplitude) in the targeted muscle on five out of 10 consecutive trials. Thresholds were defined over both motor cortices in Experiments 1 and 2. Averaged over participants, the mean stimulation levels in Experiment 1 were 44.3 (SE = 1.9) and 42.7% (SE = 1.6) of maximum stimulator output (MSO), for left and right M1, respectively. In Experiment 2, the corresponding values were 41.4 (SE = 1.1) and 42.8% (SE = 1.5) of MSO. Only left M1 was tested in Experiment 3 (mean stimulation level = 41.0% , SE = 2.2). The intensity of TMS during the experimental session was set to 115% of rMT on an individual basis.

EMG activity was recorded from surface electrodes placed over the FDI and ADM muscles of the left and right hands. EMG data were collected for 3 sec during each trial, beginning 250 ms before the onset of the fixation marker. The EMG signals were amplified (Delsys, Inc.) and bandpass filtered on-line (50-2000 Hz). The signals were analyzed offline with customized Matlab software to measure peak-to-peak amplitudes of the MEPs. Trials with background EMG activity > 20 μV in the 200 ms window preceding the TMS pulse were excluded from the analysis to prevent contamination of the MEP measurements from movement-related background EMG (Duque et al., 2005, 2007). Less than 2% of all of the trials were eliminated based on this criterion.

The primary analysis in this study focuses on whether the MEPs, measured during the baseline period prior to the onset of the trial, can predict whether or not a response can be successfully aborted following a stop signal. To this end, we sorted the MEP data into three bins based on the outcome of the trial: Go trials, Successful Stop trials, and Failed Stop trials. We then compared the MEPs for Successful Stop and Failed Stop trials, with each category compared to Go trials. We predicted that the baseline MEPs would be relatively larger prior to Failed Stop trials, based on the idea that an increased state of excitability in the motor pathways would make it more difficult to abort a subsequent response.

Given that responses are made with one of two effectors, we can also ask if the relationship between baseline excitability and stop trial outcome is limited to the probed effector, or whether effects of baseline excitability are more generic. To answer this question, we classified the muscles based on their functional role. In Experiment 1, a muscle was *task-relevant* when it was part of the response set (FDI and ADM of the responding hand). Within this category, a muscle was a *selected* respondent when the cue indicated that the muscle would be the agonist for the forthcoming response (e.g., FDI for left side cue). A muscle was classified as the *non-selected* respondent when it was in the response set, but not selected on that particular trial (e.g., ADM for left side cue). Muscles that were not part of the response set were classified as *task-irrelevant, and could either be homologous* or *non-homologous*, based on their relationship to the

selected respondent. In Experiment 2, the choice task was between the index fingers, and we only focused on task-relevant muscles (left and right FDI). In Experiment 3, we focus on right FDI only, classified the role of this muscle as either the selected or non-selected respondent.

We also sorted the data to ask if the behavior on a trial produced a change in the excitability state of the corticospinal pathway that persisted across the inter-trial interval leading up to the next trial. To answer this question, we reclassified the data as a function of the behavioral outcome of the previous trial (Go, Successful Stop, Failed Stop). MEPs were again examined based on the muscle's functional role.

3.3 Results

3.3.1 Experiment 1

The means for Go RTs, Failed Stop RTs, the estimated SSRTs, and the percentage of successful stops are presented in Table 3.1. These results are comparable to those reported in previous stop-signal studies (e.g., Logan et al., 1984). The tracking procedure yielded stopping rates of approximately 50%.

	Practice		TMS			
			Left M1 Stimulation		Right M1 Stimulation	
	R Index	R Pinkie	R Index	R Pinkie	R Index	R Pinkie
Go RT (ms)	389 (19.03)	369 (21.94)	365 (20.10)	348 (22.08)	360 (19.61)	331 (21.32)
Failed Stop RT (ms)	329 (13.08)	330 (16.71)	326 (17.27)	309 (17.20)	318 (16.52)	303 (14.68)
SSD (ms)	144 (16.62)	133 (9.93)	109 (17.83)	107 (9.28)	109 (17.83)	107 (9.28)
SSRT (ms)	245 (12.63)	237 (13.90)	256 (14.95)	241 (15.67)	251 (11.29)	224 (16.82)
Percentage successful stop	0.41 (0.03)	0.40 (0.03)	0.52 (0.05)	0.43 (0.07)	0.48 (0.05)	0.38 (0.08)

Table 3.1. Behavioral results for Experiment 1 (within-hand). Mean (SE) RT for Go and Failed Stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop.

The RT data were analyzed with a three-way repeated-measures ANOVA with the factors RT-type (Go vs. Failed Stop), session (practice vs. TMS), and finger (index vs. pinky). As observed in typical stop-signal tasks, RTs on Failed Stop trials were faster than those observed on Go trials, ($F(1,7) = 36.67$, $p < 0.005$). There was a marginally reliable effect of finger ($F(1,7) = 4.80$, $p = 0.065$), and this factor interacted with RT-type ($F(1,7) = 12.14$, $p < 0.05$). Pinky responses were faster than index finger responses, but only on Go trials. Importantly, there were

no reliable differences between the practice and TMS blocks ($F(1,7) = 3.62$, $p = 0.10$), indicating that the TMS pulses did not produce a reliable change in RT, nor did this factor interact with RT-type or finger. A two-way repeated-measures ANOVA (TMS side × finger) on the Go RTs was also not significant ($F(1,7) = 1.38$, $p = 0.28$). RTs were similar for trials in which the TMS pulse was either contralateral or ipsilateral to the response hand.

The SSRT remained relatively constant across conditions. There was no effect of session (practice vs. TMS: $F(1,7) = 0.07$, $p = 0.80$), indicating that TMS did not affect the time required to respond to the stop signal. The absence of an effect of TMS on SSRT was also confirmed in a repeated-measures ANOVA with TMS-side and finger as the two factors ($F(1,7) = 1.38$, $p = 0.28$).

The mean MEP amplitudes recorded from left and right FDI and ADM are presented in Table 3.2. The data were classified on the basis of the behavioral outcome (Go, Failed Stop, Successful Stop), and the functional role of the muscle (selected, non-selected, irrelevant). Note that in all conditions, FDI MEP values were considerably larger (and more reliable) than those from ADM since we positioned the coil to maximize the MEPs from FDI.

		Right (Task-relevant Hand)		Left (Resting Hand)	
		FDI	ADM	FDI	ADM
Selected (Right hand) / Homologous (Left hand)	Go (n = 8)	1.27 (0.21)	0.33 (0.09)	0.65 (0.19)	0.34 (0.10)
	Failed Stop (n = 8)	1.46 (0.24)	0.35 (0.08)	0.66 (0.19)	0.38 (0.11)
	Successful Stop (n = 8)	1.14 (0.18)	0.40 (0.12)	0.74 (0.22)	0.46 (0.20)
Non-selected (Right hand) / non-homologous (Left hand)	Go (n = 8)	1.32 (0.22)	0.33 (0.08)	0.66 (0.21)	0.34 (0.10)
	Failed Stop (n = 8)	1.30 (0.23)	0.34 (0.06)	0.65 (0.17)	0.36 (0.12)
	Successful Stop (n = 7)	1.15 (0.17)	0.34 (0.11)	0.43 (0.07)	0.33 (0.09)

Table 3.2. Mean (SE) amplitude (in millivolts) of MEPs recorded from each muscle in before each type of behavior in Experiment 1.

We first ask if MEPs, recorded just before trial onset, are related to success in responding to a subsequent stop signal. Figure 3.2A-D shows the baseline MEP values on Successful Stop and Failed Stop trials, relative to the MEP values on Go trials. When the right index finger was the selected response, MEPs from the right FDI were significantly lower prior to Successful Stop trials compared to Go trials ($t(7) = -2.40$, $p < 0.05$), and to Failed Stop trials ($t(7) = -3.56$, $p < 0.01$). This pattern is consistent with the hypothesis that a lower state of motor excitability at the onset of a trial can help participants abort a planned response. This modulating effect, however, was only limited to the selected muscle; none of the other comparisons in other muscle groups showed a significant modulation between the three trial types ($p > 0.1$). The absence of a

modulatory effect when the right ADM was selected is likely due to a lack of sensitivity for this muscle, given the non-optimal placement of the TMS coil for eliciting MEPs in this muscle.

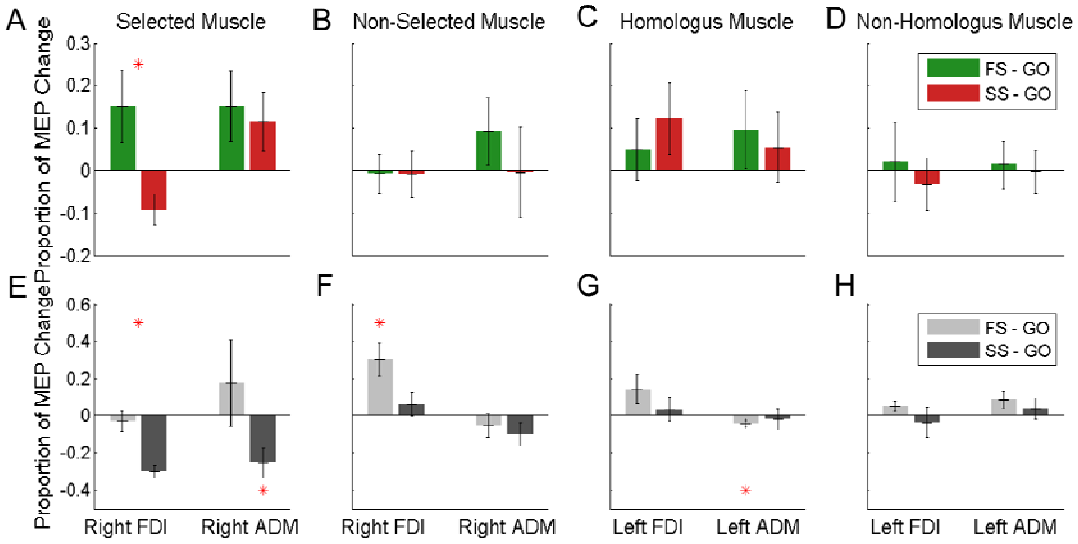

Figure 3.2. Proportion change in the amplitude of MEPs in four functional classes of muscles in Experiment 1. A-D. Proportion of MEP change *preceding* Failed Stop (FS) and Successful Stop (SS) compared to Go trials. E-F. Proportion of MEP change *after* Failed Stop (FS) and Successful Stop (SS) compared to Go trials.

We can also ask if the trial outcome produced an effect on motor excitability that lasted until the start of the subsequent trial. For example, if a trial had contained a stop signal, would the inhibitory signal generated in response to that stimulus carry over to the next trial? Figure 2E-H shows the MEP data with the trials now classified according to the outcome of the previous trial. In the selected muscle, MEPs were significantly reduced after Successful Stops compared to Go trials in both right FDI ($t(7) = -7.80$, $p < 0.001$) and ADM ($t(7) = -2.72$, $p < 0.05$). A similar effect was observed after Failed Stops in the homologous left ADM ($t(7) = -2.43$, $p < 0.05$). MEPs after Failed Stops were significantly increased compared to Go trials in the nonselected right FDI ($t(7) = 3.18$, $p < 0.05$). These findings suggest that there was indeed a carryover effect in motor excitability from the previous behavioral outcome that lasted over the inter-trial interval. The reduction of MEPs after stop trials was pronounced in the selected muscle, and also evident in the task-irrelevant, homologous muscle. The increase of MEPs in the task-relevant, non-selected muscle suggests that the consequence of an inhibitory signal is effector specific rather than generic. Inhibition directed to a selected muscle may reduce inhibition directed at a non-selected muscle. On the other hand, this increase of MEPs may also suggest that the intrinsic fluctuation being at a higher state. This is the case when one failed to issue, rather than a weaker inhibition process. The fact that we did observe the fluctuation modulation in the trial-n analysis speaks to the second interpretation.

3.3.2 Experiment 2

In Experiment 2, the choice task was between the left and right index fingers. The data from one participant was excluded because he rarely stopped his right hand during the TMS blocks (< 0.6%). The behavioral data are presented in Table 3.3. The RT data showed significant main effects of RT-type ($F(1,13) = 35.14$, $p < 0.001$) and finger ($F(1,13) = 21.31$, $p < 0.001$), with RTs faster on Failed Stop trials and faster when the responses were made with the right hand. RTs were also faster during the TMS blocks ($F(1,13) = 8.42$, $p < 0.05$). The side of stimulation (contralateral or ipsilateral to the response hand) was not reliable ($F(1,13) = 2.81$, $p = 0.12$). As in Experiment 1, SSRT remained relatively constant over the course of the experiment (no effect of session, finger, or TMS-side).

	Practice		TMS			
			Left M1 Stimulation		Right M1 Stimulation	
	L Index	R Index	L Index	R Index	L Index	R Index
Go RT (ms)	421 (24.45)	416 (25.65)	409 (22.87)	387 (23.46)	394 (18.38)	372 (21.27)
Failed Stop RT (ms)	383 (21.63)	351 (15.46)	360 (18.57)	334 (14.98)	343 (13.82)	329 (15.38)
SSD (ms)	182 (26.96)	166 (23.36)	167 (25.11)	155 (27.08)	167 (25.11)	155 (27.08)
SSRT (ms)	238 (15.86)	250 (13.03)	242 (24.14)	232 (23.69)	227 (18.15)	217 (20.67)
Percentage successful stop	0.44 (0.02)	0.42 (.03)	0.43 (0.03)	0.42 (0.04)	0.40 (0.03)	0.37 (0.04)

Table 3.3. Behavioral results for Experiment 2 (between-hand). Mean (SE) RT for Go and Failed Stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop.

		Left FDI	Right FDI
Selected	Go (n = 8)	1.00 (0.23)	1.24 (0.26)
	Failed Stop (n = 8)	1.04 (0.24)	1.27 (0.29)
	Successful Stop (n = 8)	.96 (0.21)	1.22 (0.22)
Non-selected	Go (n = 8)	.98 (0.23)	1.19 (0.23)
	Failed Stop (n = 8)	1.04 (0.24)	1.18 (0.23)
	Successful Stop (n = 7)	1.18 (0.35)	1.29 (0.26)

Table 3.4. Mean (SE) amplitude (in millivolts) of MEPs recorded from each muscle in before each type of behavior in Experiment 2.

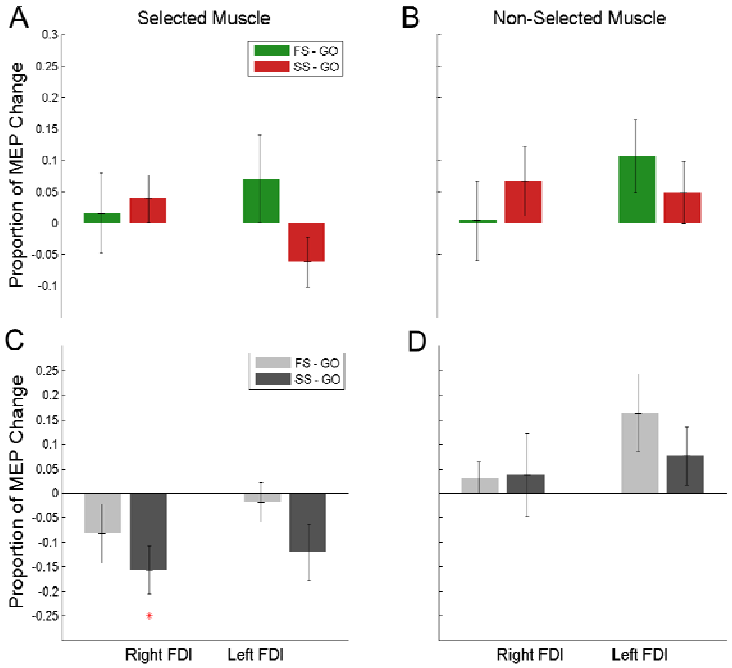

Figure 3.3. Proportion change in the amplitude of MEPs in two functional classes of muscles in Experiment 2. A-B. Proportion of MEP change *preceding* Failed Stop (FS) and Successful Stop (SS)compared to Go trials. C-D. Proportion of MEP change *after* Failed Stop (FS) and Successful Stop (SS) compared to Go trials.

Unlike Experiment 1, the state of cortical excitability prior to the onset of the trial was unrelated to the behavioral outcome on the forthcoming trial (Table 3.4, Figure 3.3A-B). MEPs prior to a Successful Stop were not smaller than those elicited prior to Go trials (right FDI: $t(13) = 1.01$, $p = 0.33$, left FDI: $t(13) = -1.47$, $p = 0.16$), or Failed Stop trials (right FDI: $t(13) = 0.24$, $p = 0.82$, left FDI: $t(13) = 0.97$, $p = 0.35$). The preceding indicates that in Experiment 2, there was no predictive effect of motor excitability on performance. However, the converse was again observed: Following a successful stop trial, MEPs in the selected muscle remained reduced at the onset of the subsequent trial, while in the non-selected muscle the direction was the opposite (Figure 3.3C-D). This effect was reliable for right FDI ($t(13) = 3.08$, $p < 0.01$) and marginally significant for left FDI ($t(13) = -2.02$, $p = 0.065$). Furthermore, MEPs were marginally larger after a Failed Stop in the non-selected left FDI ($t(13) = 2.02$, $p = 0.06$), but this effect was not observed in the right FDI ($t(13) = 0.92$, $p = 0.37$). ANOVA with finger (left vs. right) × muscle function (selected vs. non-selected) × behavioral outcome (Go vs. Failed Stop vs. Successful Stop) confirmed this observation with a significant interaction of muscle function and behavioral outcome ($F(2,24) = 3.91$, $p < 0.05$). These findings replicate, to some degree, the two kinds of carry-over effects observed in Experiment 1. First, when participants successfully generate a stop process, inhibition of cortical excitability persists for an extended period, but only in the

muscle that had been the target of that inhibition. Second, the excitability of a non-selected effector is enhanced. Both of these results indicate that the inhibitory signals are generated in an effector-specific manner. And similar to Experiment 1, the increase of MEPs in the non-selected muscle could be interpreted either as the release of the control signal when it is being directed to the selected muscle or a higher state of the non-selected muscle itself, however, in this case the absent of intrinsic fluctuation in the first place rules out the first interpretation.

3.3.3 Experiment 3

The results of Experiments 1 and 2 were inconsistent, at least with respect to how intrinsic variation in motor excitability influences subsequent task performance. Experiment 1 indicated that participants were more successful in aborting a planned response when excitability was low at the start of the trial; this effect was not observed in Experiment 2. The main methodological difference between the two studies was that Experiment 1 involved a within-hand choice and Experiment 2 a between-hand choice. Before concluding that this discrepancy is related to this methodological difference, we seek a replication in a single set of participants. Thus, in Experiment 3, we used a within-subject design, with participants completing both the within- and between-hand conditions. TMS pulses were only applied over left M1, allowing us to focus on the right FDI.

	within-hand				between-hand			
	R Index		R Pinkie		L Index		R Index	
	Practice	TMS	Practice	TMS	Practice	TMS	Practice	TMS
Go RT (ms)	384 (15.15)	381 (17.41)	380 (17.21)	371 (12.35)	407 (19.97)	386 (23.16)	398 (20.13)	382 (25.68)
Failed Stop RT (ms)	344 (10.53)	344 (12.67)	331 (12.72)	332 (11.82)	353 (14.31)	343 (19.33)	342 (13.63)	340 (20.19)
SSD (ms)	151 (12.38)	143 (17.91)	136 (9.79)	126 (12.90)	153 (9.93)	138 (16.71)	157 (13.37)	144 (18.71)
SSRT (ms)	234 (11.75)	238 (13.36)	245 (12.05)	245 (12.08)	255 (12.65)	248 (11.53)	241 (11.40)	238 (13.69)
Percentage successful stop	0.44 (0.02)	0.48 (0.04)	0.42 (0.03)	0.46 (0.03)	0.46 (0.02)	0.44 (0.05)	0.45 (0.03)	0.39 (0.04)

Table 3.5. Behavioral results for Experiment 3. Mean (SE) RT for Go and Failed Stop trials, stop-signal delay (SSD), stop-signal reaction time (SSRT), and percentage successful stop.

RTs were faster on Failed Stop trials compared to Go trials (within-hand: $F(1,13) = 52.36$, $p < 0.001$; between-hand: $F(1,13) = 50.62$, $p < 0.001$). Responses were marginally faster on right index finger trials in the between-hand comparison ($F(1,13) = 4.19$, $p = 0.065$), but there were no differences between the index and pinky fingers in the within-hand comparison ($F(1,13) = 1.52$,

p = 0.24). TMS again did not affect RT (within-hand: (F(1,13) = 1.31, p = 0.28; between-hand: F(1,13) = 0.39, p = 0.55), and SSRT was comparable across conditions.

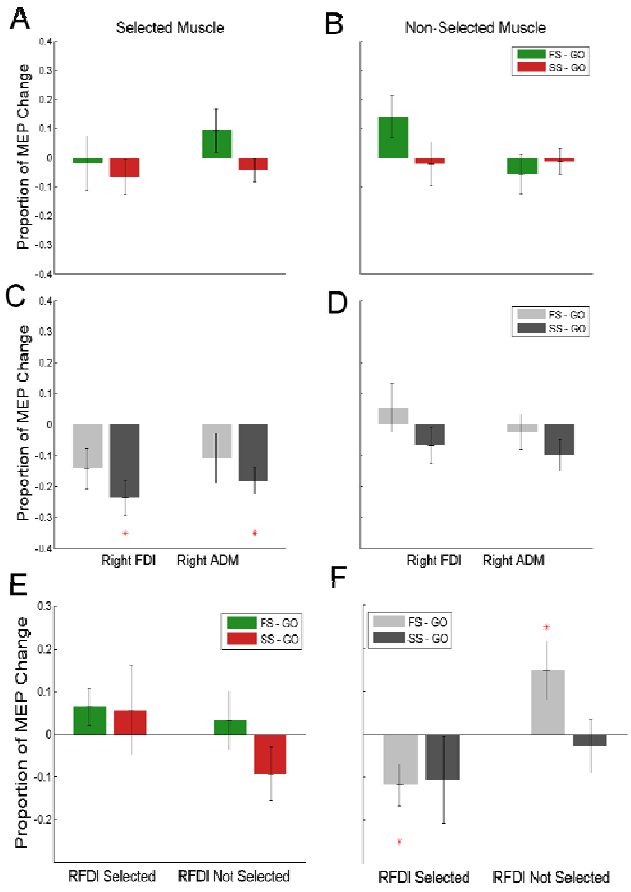

Figure 3.4. Proportion change in the amplitude of MEPs in four functional classes of muscles in Experiment 3. A-D. Data for within-hand task: proportion of MEP change *preceding* Failed Stop (FS) and Successful Stop (SS) compared to Go trials (A & B), and proportion of MEP change *after* Failed Stop (FS) and Successful Stop (SS) compared to Go trials (C & D). E-F. Data for between-hand task: proportion of MEP change *preceding* Failed Stop (FS) and Successful Stop (SS) compared to Go trials (E), and proportion of MEP change *after* Failed Stop (FS) and Successful Stop (SS) compared to Go trials (F).

The MEP data are presented in Table 3.6. We did not observe any relationship between the state of cortical excitability prior to trial onset on subsequent task performance. This null result was found for the between-hand comparison between MEPs prior to Successful Stop and Failed Stop (FDI: t(13) = 0.01, p = 0.99), consistent with the results of Experiment 2. Moreover, it was

also not found in the within-hand comparison (FDI: $t(13) = 0.42$, $p = 0.68$, ADM: $t(13) = -1.07$, $p = 0.31$), constituting a failure to replicate the results of Experiment 1.

Carry-over effects were reliable in Experiment 3. In the within-hand task, cortical excitability in right FDI was reduced after Successful Stop trials compared to Go trials when the right index finger had been selected (FDI: $t(11) = -3.93$, $p < 0.005$, ADM: $t(11) = -4.09$, $p < 0.005$). There was also a trend for reduced excitability after Failed Stop trials ($t(11) = -2.06$, $p = 0.06$). In the between-hand task, the lingering inhibition of the selected finger was only reliable after Failed Stops ($t(11) = -2.38$, $p < 0.05$) (after Successful Stop: $t(11) = -1.03$, $p = 0.33$). Same as in Experiment 2, the interaction between muscle function and behavior was significant ($F(2,20) = 4.52$, $p < 0.05$), indicating a trend of reduction of MEPs in the selected muscle, while an increase of MEPs in the non-selected muscle after Stop trials. In sum, without the intrinsic fluctuation modulation in the trial-n analysis, these results more likely indicate that following the generation of a stop command, inhibition persists across the inter-trial interval, but only in the targeted muscle.

		With-in Hand Task		Between-hand task
		R FDI	R ADM	L FDI
Selected (Right hand) / Homologous (Left hand)	Go (n = 8)	0.71 (0.12)	0.25 (0.06)	1.07 (0.16)
	Failed Stop (n = 8)	0.63 (0.08)	0.27 (0.07)	1.14 (0.18)
	Successful Stop (n = 8)	0.67 (0.12)	0.24 (0.06)	1.06 (0.15)
Non-selected (Right hand) / non-homologous (Left hand)	Go (n = 8)	0.69 (0.11)	0.25 (0.06)	1.09 (0.16)
	Failed Stop (n = 8)	0.77 (0.11)	0.23 (0.06)	1.10 (0.16)
	Successful Stop (n = 7)	0.65 (0.11)	0.25 (0.06)	1.02 (0.18)

Table 3.6. Mean (SE) amplitude (in millivolts) of MEPs recorded from each muscle in before each type of behavior in Experiment 3.

3.4 Discussion

Results in the three experiments were mixed. Experiment 1 showed that in selected muscles, fluctuation in the MEPs modulates the successfulness of stopping. This effect was effector specific, and was not observed in the other categories of muscles. In Experiment 2, no modulation was observed in a between-hand task, neither in Experiment 3, with the within-subject design of both unimanual and bimanual stopping. The carry-over effects of inhibition vs. heightened state from the previous trial may address these apparent inconsistent findings. In all three experiments, we observed a trend of reduction of MEPs after Successful Stops in the selected muscle and an increase of MEPs after Failed Stops in the non-selected muscle. The

carry-over effect in the Failed Stops in non-selected muscles could be attributed to a release of inhibition signal issued via control mechanism or in general a heightened state in these muscles. The absence of the intrinsic fluctuation modulation in Experiment 2 and 3 disconfirms the first interpretation.

3.4.1 Interplay of the extrinsic and intrinsic activities: the case of failed stop

When encountering a stop-signal, inhibition was carried over to the ITI period, resulting in reduced CS excitabilities after Successful Stops. During the Failed Stop, the control signal and the internal state "carried over" are pointing to the opposite directions in the selected muscle, while in the non-selected muscle they are towards the same direction, resulting in increased MEPs. However, depending on at which stage the failure occurs, the elevation of CS excitability could be due to the control signal or the intrinsic fluctuation. At an early stage, a failure of responding to the stop signal could be due to some attentional failure. An increased MEP after Failed Stop therefore would be a carry-over of a higher intrinsic state. At a later stage when the inhibition process failed to countermand the go process, directing the inhibition to the selected muscle may result in the release of inhibition in the non-selected muscle, thus the increased MEPs. The two interpretations could be teased apart by the trial-n observations, as well as the MEP levels in the selected muscle during Failed Stops. In both Experiment 2 and 3, there was a trend of MEP reduction in the selected muscles no matter the stop was successful or failed. Also, there was no intrinsic fluctuation modulation observed in trial-n analysis. Both add to the evidence supporting the carry-over of control process hypothesis. However, in Experiment 1, the absence of reduction in MEPs after Failed Stop, plus the modulation effect of intrinsic fluctuation, supports the hypothesis that the higher internal state was carried over after Failed Stops.

It is, however, not easy to establish the real cause-effect relations of the extrinsic and intrinsic processes and the behavior outcome. Nor it is clear how they interact. Do they simply superimpose upon each other or does one re-organize another? These are questions for further investigation.

3.4.2 Sequential effects in the stop-signal task

Many stop-signal studies found that Go RTs slow down after encountering a stop signal in the previous trial, especially after Successful Stop (Rieger and Gauggel, 1999, Verbruggen et al., 2008). This after effect has been attributed to shifting response criterion, between trial control processes or repetition priming (Rieger and Gauggel, 1999), and enhanced association between stop signal and inhibition and memory retrieval (Verbruggen et al., 2008). In our experiments, the consistent finding is that MEPs in the selected muscle were decreased after Successful Stops.

Our research question of the influence of intrinsic fluctuation offers a different angle of looking at this effect. Failed Stops often register as an error in the system. We should expect that the memory trace would be even stronger. The consistent finding is that in the between-hand tasks, MEPs in the selected muscles decreased after Failed Stop. Also, as there was no evidence of intrinsic fluctuation modulation in these experiments, the finding is consistent with the hypothesis that the control processes was carried over to the beginning of the next trial. While in the within-hand task, the finding is mixed. In Experiment 1, there was no change in selected-muscle MEPs after Failed Stops, whereas there was a trend of reduction. This would be hard to explain with traditional approach focusing on RT analysis. Here the finding in the intrinsic fluctuation modulation provides us additional information. In Experiment 1, there was a modulation effect, thus the null effect in selected muscle after Failed Stop is more likely to be a mixture of extrinsic control process and intrinsic fluctuation; while in Experiment 3, the absence of intrinsic fluctuation modulation speaks to a stronger inhibition signal, which is consistent with the trend of reduced MEPs after both Successful and Failed Stops.

3.4.3 Neural substrates of intrinsic fluctuation

Using a repeated movie presentation paradigm, Golland et al. (2006) found two networks in the posterior cortex: the extrinsic network that is correlated with task-related activation, and an intrinsic network that is less correlated with the external stimuli. These networks may be candidates for neural substrats of the externally- and internally-oriented processes.

Another question is what the frequencies of the intrinsic fluctuations are. Gamma frequency band seem to be a candidate. Simultaneous recordings of EEG and fMRI showed correlation between the spontaneous fluctuation in BOLD signals and the gamma band activities (Laufs, 2003).

3.4.4 Conclusion

In summary, the results from the three experiments in the current study showed an interaction between the intrinsic fluctuation and extrinsic control process in inhibitory control. The control process may countermand the fluctuation, depending on the strength of both processes. This can be teased apart by looking at the "carry-over" effect of a successful or failed stop. When inhibition is strong, as evident in the between-hand stop-signal task, it is carried over to the next trial; while when the control process is weaker, as shown in the within-hand task, the next trial also exhibits some modulation of fluctuation on the successfulness of stopping.

Chapter 4

A Rational Analysis of Serial Reproduction

4.1 Introduction

At the heart of our analysis of serial reproduction is the question of how memory biases influence cultural transmission. Biased reconstructions are found in many tasks. For example, people are biased by their knowledge of the structure of categories when they reconstruct simple stimuli from memory. One common effect of this kind is that people judge stimuli that cross boundaries of two different categories to be further apart than those within the same category even when the distances between the stimuli are the same (Liberman, Cooper, shankweiler, & Studdert-Kennedy, 1967). However, biases need not reflect suboptimal performance. If we assume that memory is solving the problem of extracting and storing information from the noisy signal presented to our senses, we can analyze the process of reconstruction from memory as a Bayesian inference. Under this view, reconstructions should combine prior knowledge about the world with the information provided by noisy stimuli. Use of prior knowledge will result in biases, but these biases ultimately make memory more accurate (Huttenlocher, Hedges, & Vevea, 2000).

If this account of reconstruction from memory is true, we would expect the same inference process to occur at every step of serial reproduction. The effects of memory biases should thus be accumulated. Assuming all participants share the same prior knowledge about the world, serial reproduction should ultimately reveal the nature of this knowledge. Drawing on recent work exploring other processes of information transmission (Griffiths & Kalish, 2005, 2007), we show that a rational analysis of serial reproduction makes exactly this prediction. To test this account, we explore the special case where the task is to reconstruct a one-dimensional stimulus using the information that it is drawn from a fixed Gaussian distribution. In this case we can precisely characterize behavior at every step of serial reproduction. Specifically, we show that this defines a simple first-order autoregressive, or AR(1), process, allowing us to draw on a variety of results characterizing such processes. We use these predictions to test Bayesian models of serial reproduction in a series of laboratory experiments, and show that the predictions hold for serial reproduction both between- and within-subjects.

The plan of the paper is as follows. We first lay out the Bayesian account of serial reproduction, starting with a rational analysis of reconstruction from memory. We then show how this Bayesian account corresponds to the AR(1) process in the case of simple Gaussian distributions, and how the AR(1) model can accommodate different assumptions about memory storage and reconstruction. In the main body of the paper, we use this model to motivate four experiments testing the prediction that serial reproduction reveals memory biases. Finally, we consider the implications of the results of these experiments in the General Discussion.

4.2 A Bayesian view of serial reproduction

We will outline our Bayesian approach to serial reproduction by first considering the problem of reconstruction from memory, and then asking what happens when the solution to this problem is repeated many times, as in serial reproduction.

4.2.1 Reconstruction from memory

Our goal is to give a rational account of reconstruction from memory, considering the underlying computational problem and finding the optimal solution to that problem. We will formulate the problem of reconstruction from memory as a problem of inferring and storing accurate information about the world from noisy sensory data. Given a noisy stimulus x, we seek to recover the true state of the world μ that generated that stimulus, storing an estimate $\hat{\mu}$ in memory. The optimal solution to this problem is provided by Bayesian statistics. Previous experience provides a "prior" distribution on possible states of the world, $p(\mu)$. On observing x, this can be updated to a "posterior" distribution $p(\mu|x)$ by applying Bayes' rule

$$p(\mu|x) = \frac{p(x|\mu)p(\mu)}{\int p(x|\mu')p(\mu')\,d\mu'} \tag{4.1}$$

where $p(x|\mu)$ – the "likelihood" – indicates the probability of observing x if μ is the true state of the world. Having computed $p(\mu|x)$, a number of schemes could be used to select an estimate of$\hat{\mu}$ to store. Perhaps the simplest such scheme is sampling from the posterior, with $\hat{\mu} \sim p(\mu|x)$.

This analysis provides a general schema for modeling reconstruction from memory, applicable for any form of x and μ. A simple example is the special case where x and μ vary along a single continuous dimension. In the experiments presented later in the paper we take this dimension to be the width of a fish, showing people a fish and asking them to reconstruct its width from memory, but the dimension of interest could be any subjective quantity such as the perceived length, loudness, duration, or brightness of a stimulus. Assume that previous experience establishes that μ has a Gaussian distribution, with $\mu \sim N(\mu_0, \sigma_0^2)$, and that the noise process means that x has a Gaussian distribution with μ as its center, $x|\mu \sim N(\mu, \sigma_x^2)$. In this case, we can use standard results from Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 1995) to show that the outcome of Equation

4.1 is also a Gaussian distribution, with $p(\mu|x)$ being $N(\lambda x + (1-\lambda)\mu_0, \lambda\sigma_x^2)$, where $\lambda = 1/(1+\sigma_x^2/\sigma_0^2)$.

The analysis presented in the previous paragraph makes a clear prediction: that the reconstruction $\hat{\mu}$ should be a compromise between the observed value x and the mean of the prior μ_0, with the terms of the compromise being set by the ratio of the noise in the data σ_x^2 to the uncertainty in the prior σ_0^2. This model thus predicts a systematic bias in reconstruction that is not a consequence of an error of memory, but the optimal solution to the problem of extracting information from a noisy stimulus. The possibility that memory biases might actually be the consequence of a process that improves the accuracy of memory was pointed out by Huttenlocher et al. (2000), who presented a model extremely similar to that outlined in this section.

Huttenlocher et al. (2000) conducted several experiments testing this account of memory biases using simple one-dimensional stimuli such as fish that vary in width. In each experiment participants were shown stimuli sampled from a probability distribution such as a Gaussian (i.e. normal distribution) and asked to reconstruct those stimuli from memory. The results showed that people's reconstructions interpolated between the observed stimuli and the mean of the trained distribution as predicted. A similar notion of reconstruction as a weighted average of the mean of prior distribution and an observation was used by Hemmer and Steyvers (2008), who found that people formed appropriate Bayesian reconstructions for realistic stimuli such as images of fruit, and seemed capable of drawing on prior knowledge at multiple levels of abstraction in doing so. Finally, Stewart, Brown, and Chater (2005) showed that a similar kind of biased estimation appears in sequential retrieval from memory.

4.2.2 Serial reproduction

With a model of how people might approach the problem of reconstruction from memory in hand, we are now in a position to analyze what happens in serial reproduction, where the stimuli that people receive on one trial are the results of a previous reconstruction. On the nth trial, a participant sees a stimulus x_n. The participant then computes $p(\mu|x_n)$ as outlined in the previous section, and stores a sample $\hat{\mu}$ from this distribution in memory. When asked to produce a reconstruction, the participant generates a new value x_{n+1} from a distribution that depends on $\hat{\mu}$. If the likelihood, $p(x|\mu)$, reflects perceptual noise, then it is reasonable to assume that x_{n+1} will be sampled from this distribution, substituting $\hat{\mu}$ for μ. This value of x_{n+1} is the stimulus for the next trial.

Viewed from this perspective, serial reproduction defines a stochastic process: a sequence of random variables evolving over time. In particular, it is a Markov chain, since the reconstruction produced on the current trial depends only on the value produced on the preceding trial (e.g. Norris, 1997). The transition probabilities of this Markov chain are

$$p(x_{n+1}|x_n) = \int p(x_{n+1}|\mu)\, p(\mu|x_n)\, d\mu \tag{4.2}$$

being the probability that x_{n+1} is produced as a reconstruction for the stimulus x_n, which is also the

reconstruction from previous trial. If this Markov chain is ergodic (see Norris, 1997 for details) it will converge to a stationary distribution $\pi(x)$, with $p(x_n|x_1)$ tending to $\pi(x_n)$ as $n \to \infty$. That is, after many reproductions, we should expect the probability of seeing a particular stimulus being produced as a reproduction to stabilize to a fixed distribution. Identifying this distribution will help us understand the consequences of serial reproduction.

The transition probabilities given in Equation 4.2 have a special form, being the result of sampling a value from the posterior distribution $p(\mu|x_n)$ and then sampling a value from the likelihood $p(x_{n+1}|\mu)$. In this case, it is possible to identify the stationary distribution of the Markov chain (Griffiths & Kalish, 2005, 2007). The stationary distribution of this Markov chain is the *prior predictive distribution*

$$\pi(x) = \int p(x|\mu)p(\mu)\,d\mu \tag{4.3}$$

being the probability of observing the stimulus x when μ is sampled from the prior. This happens because this Markov chain is a *Gibbs sampler* for the joint distribution on x and μ defined by multiplying $p(x|\mu)$ and $p(\mu)$ (Griffiths & Kalish, 2007). A Gibbs sampler is a Markov chain defined by alternating between sampling from the conditional distributions $p(x|\mu)$ and $p(\mu|x)$ for some joint distribution $p(x,\mu)$, which results in $p(x,\mu)$ as a stationary distribution. This gives a clear characterization of the consequences of serial reproduction: after many reproductions, the stimuli being produced will be sampled from the prior distribution assumed by the participants. Convergence to the prior predictive distribution provides a formal justification for the traditional claims that serial reproduction reveals cultural biases (e.g., Bartlett, 1932), since those biases would be reflected in the prior.

The convergence results given in the previous paragraph are completely general, applying to any kind of stimuli, hypotheses, and prior. In the special case of reconstruction of stimuli that vary along a single dimension, we can also analytically compute the probability density functions for the transition probabilities and stationary distribution. Applying Equation 4.2 using the results summarized in the previous section, we have $x_{n+1}|x_n \sim N(\mu_n, (\sigma_x^2 + \sigma_n^2))$, where $\mu_n = \lambda x_n + (1-\lambda)\mu_0$, and $\sigma_n^2 = \lambda\sigma_x^2$. Likewise, Equation 4.3 indicates that the stationary distribution is $N(\mu_0, (\sigma_x^2 + \sigma_0^2))$. The rate at which the Markov chain converges to the stationary distribution depends on the value of λ. When λ is close to 1, convergence is slow since μ_n is close to x_n. As λ gets closer to 0, μ_n is more influenced by μ_0 and convergence is faster. Since $\lambda = 1/(1+\sigma_x^2/\sigma_0^2)$, the convergence rate thus depends on the ratio of the participant's perceptual noise and the variance of the prior distribution, σ_x^2/σ_0^2. More perceptual noise results in faster convergence, since the specific value of x_n is trusted less; while more uncertainty in the prior results in slower convergence, since x_n is given greater weight.

4.3 Serial reproduction of one-dimensional stimuli as autoregression

The special case of serial reproduction of one-dimensional stimuli can also give us further insight into the consequences of modifying our assumptions about storage and reconstruction from memory, by exploiting a further property of the underlying stochastic process: that it is a first-order autoregressive process, abbreviated to AR(1). The general form of an AR(1) process is

$$x_{n+1} = c + \phi x_n + \varepsilon_{n+1} \tag{4.4}$$

where $\varepsilon_{n+1} \sim N(0, \sigma_\varepsilon^2)$. Equation 4.4 has the familiar form of a regression equation, predicting one variable as a linear function of another, plus Gaussian noise. It defines a stochastic process because each variable is being predicted from that which precedes it in sequence. AR(1) models are widely used to model timeseries data, being one of the simplest models for capturing temporal dependency (Box & Jenkins, 1994).

Just as showing that a stochastic process is a Markov chain provides information about its dynamics and asymptotic behavior, showing that it reduces to an AR(1) process provides access to a number of results characterizing the properties of these processes. If $\phi < 1$ the process has a stationary distribution that is Gaussian with mean $c/(1-\phi)$ and variance $\sigma_\varepsilon^2/(1-\phi^2)$. The autocovariance at a lag of n iterations is $\phi^n \sigma_\varepsilon^2/(1-\phi^2)$, and thus decays geometrically in ϕ. An AR(1) process thus converges to its stationary distribution at a rate determined by ϕ.

It is straightforward to show that the stochastic process defined by serial reproduction where a sample from the posterior distribution on μ is stored in memory and a new value x is sampled from the likelihood is an AR(1) process. Using the results in the previous section, at the $(n+1)$th iteration

$$x_{n+1} = (1-\lambda)\mu_0 + \lambda x_n + \varepsilon_{n+1} \tag{4.5}$$

where $\lambda = 1/(1+\sigma_x^2/\sigma_0^2)$ and $\varepsilon_{n+1} \sim N(0, (\sigma_x^2 + \sigma_n^2))$ with $\sigma_n^2 = \lambda\sigma_x^2$. This is an AR(1) process with $c = (1-\lambda)\mu_0$, $\phi = \lambda$, and $\sigma_\varepsilon^2 = \sigma_x^2 + \sigma_n^2$. Since λ is less than 1 for any σ_0^2 and σ_x^2, we can find the stationary distribution by substituting these values into the expressions given above. As described in the previous section, the value of λ determines the trade-off between the prior and the current piece of data, and thus the convergence rate of the Markov chain.

Identifying serial reproduction for single-dimensional stimuli as an AR(1) process allows us to relax our assumptions about the way that people are storing and reconstructing information. In the memorization phase, the participant's memory $\hat{\mu}$ can be 1) a sample from the posterior distribution $p(\mu|x_n)$, as assumed above, or 2) a value such that $\hat{\mu} = \text{argmax}_\mu\, p(\mu|x_n)$, being the posterior mode, which is also the mean of the posterior since the mode of a Gaussian is its mean. In the reproduction phase, the participant's reproduction x_{n+1} can be 1) a noisy reconstruction, which is a sample from the likelihood $p(x_{n+1}|\hat{\mu})$, as assumed above, or 2) a perfect reconstruction from memory, such that $x_{n+1} = \hat{\mu}$. This defines four different models of serial reproduction:

1. *Sample - Sample (SS)*: participants store a sample from posterior distribution of their noisy observation, and their reconstruction is also a sample from memory
2. *Sample - Perfect (SP)*: participants store a sample from posterior distribution of their noisy observation, and they give a perfect reconstruction from memory
3. *Maximize - Sample (MS)*: participants store the expected value of the posterior distribution of their noisy observation, and their reconstruction is also a sample from memory
4. *Maximize - Perfect (MP)*: participants store the expected value of the posterior distribution of their noisy observation, and they give a perfect reconstruction from memory

The fourth model is unlikely to work well for human behavior, because no variance is allowed in the system, so the chain will move to the mean of the prior in a strictly monotonic fashion and stay there forever. We thus consider only the first three models in the following analysis.

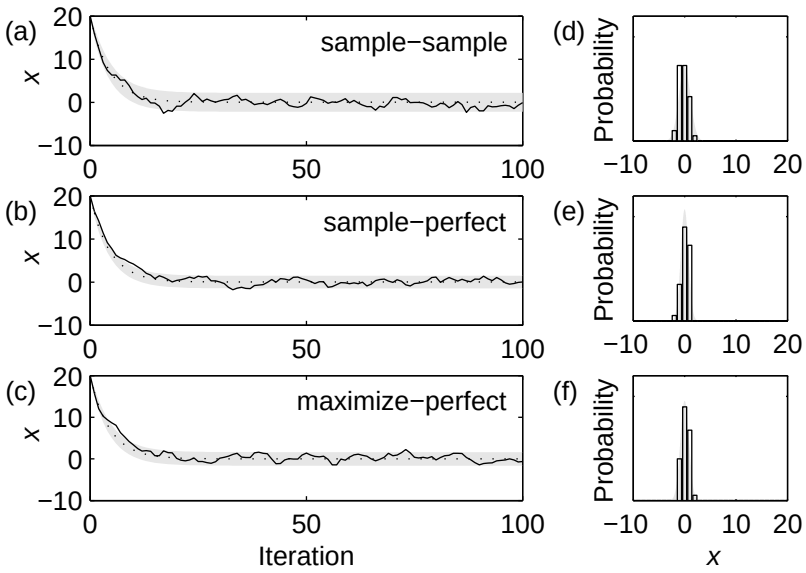

Figure 4.1: Model simulations for three types of behavior. Panels (a)-(c) show simulated Markov chains with $\sigma_x = 0.5$ and $\sigma_0 = 1.0$. The solid line in each graph is a sequence of sampled values of x_n, and the dotted line and the gray area show the mean and 95% confidence interval on x_n. All samples and statistics are conditioned on $x_1 = 20$. Panels (d)-(f) show a histogram of the last 80 values of x_n for the Markov chains in panels (a)-(c). The gray areas are the probability density functions for the stationary distributions of the three models.

All three of these models can be shown to have the general form of the AR(1) process (Equation 4.5). What distinguishes between them is the noise terms: in the SS model $\varepsilon_{n+1} \sim N(\mu_n, (\sigma_x^2 + \sigma_n^2))$;

in the SP model $\varepsilon_{n+1} \sim N(\mu_n, \sigma_n^2)$, since there is no reproduction noise; and in the MS model, $\varepsilon_{n+1} \sim N(\mu_n, \sigma_x^2)$ as no noise is introduced in memory storage. As shown above, the SS model converges to the prior predictive distribution $N(\mu_0, (\sigma_x^2 + \sigma_0^2))$. Using the autoregression analysis, it is easy to show that the SP and MS models converge to $N(\mu_0, \frac{\lambda\sigma_x^2}{1-\lambda^2})$ and $N(\mu_0, \frac{\sigma_x^2}{1-\lambda^2})$ respectively. Thus, all three models converge to a distribution determined by the prior. Figure 4.1 shows a simulation of the three types of models. The stationary distribution in the SS model has the largest variance and that of the SP model has the smallest variance when $\sigma_x^2 < \sigma_0^2$.

More generally, any model in which reconstructions have a mean value corresponding to $(1-\lambda)\mu_0 + \lambda x_n$ and storage and reconstruction are subject to Gaussian noise will reduce to an AR(1) process, with the only variation between models appearing in the variance of the noise term ε. This significantly increases the generality of our characterization of serial reproduction, but it also means that different models developed within this framework (including the three models introduced above) cannot be differentiated using empirical data. However, all of these models make the same basic prediction: that repeatedly reconstructing stimuli from memory will result in convergence to a distribution whose mean corresponds to the mean of the prior.

4.3.1 Testing the model predictions

In the remainder of the paper we describe four experiments testing the predictions produced by this model. These experiments all use a serial reproduction paradigm with stimuli that vary only along one dimension (the width of fish, following Huttenlocher et al. 2000). By using these simple stimuli, we are able to provide the first carefully-controlled empirical analysis of the effects of serial reproduction, and test whether the results match the quantitative predictions produced by our model. Huttenlocher et al. (2000) established that people behave in a way that is consistent with the Bayesian analysis of reconstruction presented above through a series of experiments using these stimuli. Our experiments thus replicate and extend these results to the case of serial reproduction.

Experiment 1 follows previous research on serial reproduction in using a between-subjects design, with the reconstructions of one participant serving as the stimuli for the next. Participants were trained on the distribution of fish widths associated with a category, establishing a prior distribution for use in reconstruction. Experiment 2 uses a within-subjects design in which each person reconstructs stimuli that they themselves produced on a previous trial, testing the potential of this design to reveal the memory biases of individuals. Experiment 3 removes the training on the prior distribution, allowing serial reproduction to reveal prior expectations derived from general world knowledge instead of laboratory training. The results of the experiment reveal a general bias that was also reflected in the results of Experiments 1 and 2, helping to explain a trend observed in the previous experiments. In Experiment 4, we explore the consequences of serial reproduction with a more complex prior – a bimodal distribution – and examine the influence of categories and context on reconstruction.

4.4 Experiment 1: Between-Subjects serial reproduction

This experiment directly tested the basic prediction that the outcome of serial reproduction will reflect the prior knowledge that people have about the distribution of stimuli. The experiment followed the same basic procedure as Bartlett's (1932) classic experiments, using the reconstruction task introduced by Huttenlocher et al. (2000). Two groups of participants were trained on different distributions of a one-dimensional quantity – the width of a schematic fish – that would serve as a prior for reconstructing similar stimuli from memory. The distributions learned by the two groups differed in their means, allowing us to examine whether the mean of the distribution produced by serial reproduction is affected by the prior, as predicted by our model.

4.4.1 Method

Participants

Forty-six undergraduates from the University of California, Berkeley participated in exchange for course credit.

Stimuli

Stimuli were the same as those used in Huttenlocher et al. (2000): fish with elliptical bodies and fan-shaped tails. All the fish stimuli varied only in one dimension, the width of the fish, ranging from 2.63cm to 5.76cm. The stimuli were presented on an Apple iMac computer by a Matlab script using PsychToolBox extensions (Brainard, 1997; Pelli, 1997).

Procedure

Participants received instructions that indicated that they would be working at a fish farm, and would receive some on-the-job training before beginning work. They were then trained to discriminate between two categories of fish: fish-farm and ocean fish. The width of the fish-farm fish was normally distributed and that of the ocean fish was uniformly distributed between 2.63 and 5.75cm. To make the training process easier, the instructions explained that fish-farm fish are fed on a special diet and are thus vary around a standard size, while ocean fish have to fend for themselves and have a far greater range of sizes. The critical manipulation was the parameters of the normal distribution, with two groups of participants being trained on distributions with different means but the same standard deviation. In condition A, $\mu_0 = 3.66$cm, $\sigma_0 = 1.3$cm; in condition B, $\mu_0 = 4.72$cm, $\sigma_0 = 1.3$cm.

In the training phase, participants first received a block of 60 trials. On each trial, a stimulus was presented at the center of a computer monitor and participants tried to predict which type of fish it was by pressing one of the keys on the keyboard and they received feedback about the correctness of the prediction. The participants were then tested for 20 trials on their knowledge of the two types of fish. The procedure was the same as the training block except there was no

feedback. The training-testing loop was repeated until the participants reached 80% of optimal performance.[1] If a participant did not reach this criterion after five iterations, the experiment halted. All the participants passed the training phase.

In the reproduction phase, the participants were told that they were going to begin to work on recording fish sizes for the fish farm. On each trial, a fish stimulus was flashed at the center of the screen for 500ms and then disappeared. Another fish of random size appeared at one of four possible positions near the center of screen and the participants used the up and down arrow keys to adjust the width of the fish until they thought it matched the fish they just saw. The fish widths seen by the first participant in each condition were 120 values uniformly spanning the range from 2.63 to 5.75cm. The first participant reconstructed these stimuli from memory. Each subsequent participant in each condition was then presented with the reconstructions produced by the previous participant as stimuli, and they again tried to reconstruct those fish widths. Thus, the data from each participant constitute one slice of time in 120 serial reproduction chains.

At the end of the experiment, the participants were given a final 50-trial test to check if their prior distributions had drifted. Since serial reproduction builds on the responses of previous participants, a single participant who is not engaged with the task can disrupt the entire experiment. Participants were thus excluded from the experiment if their data failed any of three conditions: 1) final testing score was less than 80% of optimal performance; 2) the difference between the reproduced value and stimulus shown was greater than the difference between the largest and the smallest stimuli in the training distribution on any trial; 3) there were no adjustments from the starting value of the fish width for more than half of the trials. If the current participant's data were rejected, the next participant would see the data generated by the previous participant. A total of 10 participants were excluded following these criteria.

4.4.2 Results and discussion

There were 18 participants in each condition, resulting in 18 generations of serial reproduction. Figure 4.2 shows the initial and final distributions of the reconstructions, together with the plots for the 120 chains in the two conditions. The initial set of stimuli in both conditions A and B were drawn from the same uniform distribution. The mean reconstructed fish widths produced by the first participants in these conditions were 4.22 and 4.21cm respectively, which were not statistically significantly different ($t(238) = 0.09$, $p = 0.93$). The histograms in the right panel show the final distributions of the reconstructions by the 18th participants in the two conditions.

[1]The optimal decision strategy used for evaluating performance was the Bayesian strategy under 0-1 loss (i.e. assuming that some reward is received for a correct answer, but no reward for an incorrect answer). This strategy corresponds to choosing the category with highest posterior probability for each stimulus. Since our normal (fish-farm fish) and uniform (ocean fish) distributions overlap, we calculated the upper and lower bounds where the posterior probability of the stimuli under the normal distribution is greater than under the uniform distribution. The optimal decision strategy is to classify the fish as ocean fish for sizes outside these boundaries and to classify those within these boundaries as fish-farm fish. Each participant's responses could then be scored for their consistency with this strategy.

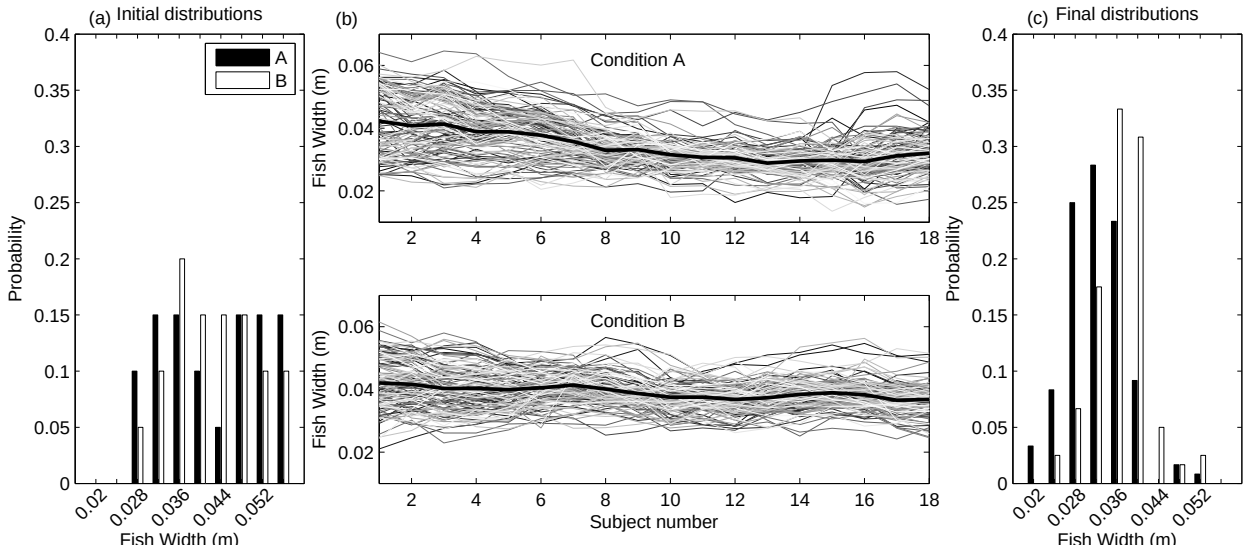

Figure 4.2: Initial and final distributions, and the chains for the two conditions in Experiment 1. (a) The histograms for the initial distributions of stimuli for conditions A and B. (b) The chains of serial reproduction, starting from the first participant's data, ending with the 18th participant's data. (c) The histograms for the ending distributions of stimuli for conditions A and B.

The mean reconstructed fish widths were 3.20 and 3.68cm respectively, a statistically significant difference ($t(238) = 6.93$, $p < 0.001$). A two-way ANOVA also showed a significant interaction between the starting and ending points of the 120 chains in the two conditions ($F(1,236) = 12.04$, $p < 0.001$). The difference in means matches the direction of the difference in the training provided in conditions A and B, although the overall size of the difference is reduced and the means of the stationary distributions were lower than those of the distributions used in training.

Figure 4.3 shows the autoregression plots and the biases in reconstruction. The autoregression plots compare x_{n+1} with x_n, and the autoregression model outlined above predicts that the resulting distribution should be jointly Gaussian with the mean of x_{n+1} being a linear function of x_n. This is exactly what we see in Figure 4.3. The correlation between the stimulus x_n and its reconstruction x_{n+1} is the correlation between the AR(1) model's predictions and the data, and this correlation was high in both conditions, being 0.91 and 0.86 (both $p < 0.001$) for conditions A and B respectively. Biases in reconstruction can be identified by evaluating the difference between x_{n+1} and x_n as a function of x_n. This was the basic dependent measure used by Huttenlocher et al. (2000). Consistent with their results, the slope of the function relating bias and x_n is negative (-0.34 for both conditions A and B, $p < 0.001$), confirming the Bayesian model's prediction that memory of the stimuli are biased towards the mean of the category distribution.

Finally, we examined whether the Markov assumption underlying our analysis was valid, by computing the correlation between x_{n+1} and x_{n-1} given x_n. The resulting partial correlation was low for both conditions, being 0.04 and 0.01 in conditions A and B respectively (both $p > 0.05$).

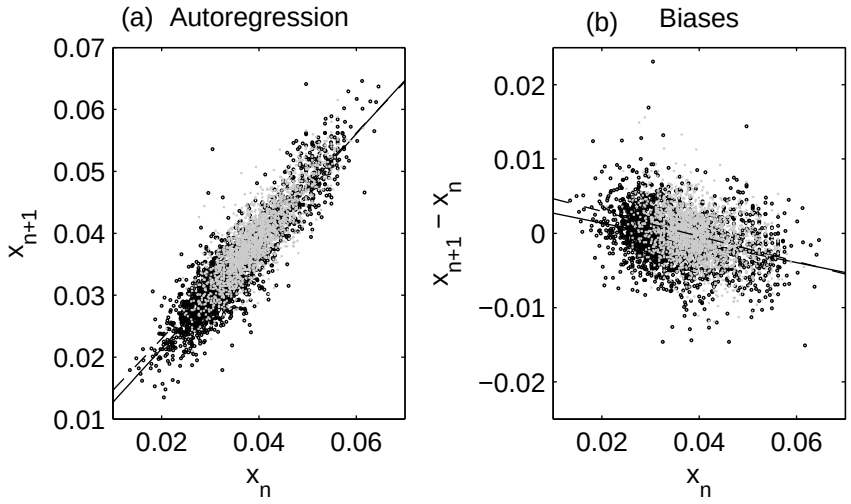

Figure 4.3: Autoregression and bias plots for the two conditions in Experiment 1. (a) The autoregression plots for conditions A (black) and B (gray). (b) The bias plots for conditions A (black) and B (gray).

This is to be expected, since the use of a different participant at each step of reproduction ensures that the Markov assumption should hold, but allows us to rule out any higher-order temporal effects on reconstruction.

The results of the experiment provide support for the basic prediction produced by our analysis of serial reproduction: chains formed of individuals trained on different prior distributions converged to different stationary distributions, and those stationary distributions differed in a way that reflected the difference in the priors. However, the stationary distributions did not correspond exactly to the prior distributions on which people were trained – a point we will return to in Experiment 3, after examining whether we obtain similar results when creating serial reproduction chains within-subjects.

4.5 Experiment 2: Within-Subjects serial reproduction

The between-subjects design allows us to reproduce the process of information transmission, but our analysis suggests that serial reproduction might also have promise as a method for investigating the memory biases of individuals. To explore the potential of this method, we tested the model with a within-subjects design, in which a participant's reproduction in the current trial became the stimulus for that same participant in a later trial. Each participant's responses over the entire experiment thus produced a chain of reproductions. Each participant produced three such chains, starting from widely separated initial values. Control trials and careful instructions were used so that the participants would not realize that some of the stimuli were their own reproductions.

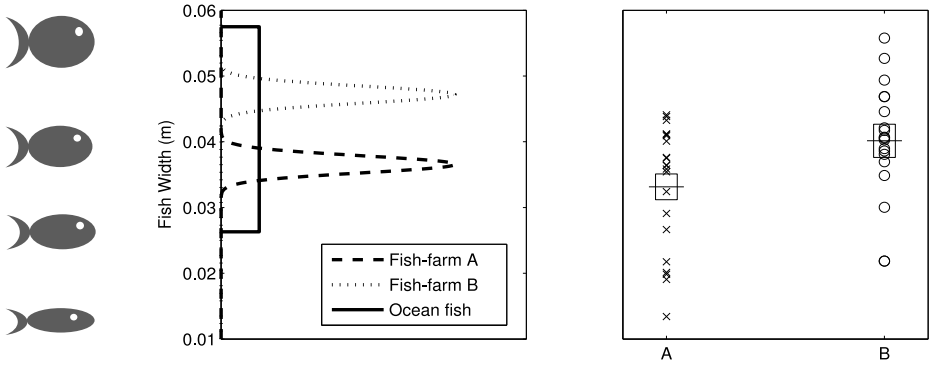

Figure 4.4: Stimuli, training distributions and stationary distributions for Experiment 2. Each data point in the right panel shows the mean of the last 20 iterations for a single participant. Boxes show the 95% confidence interval around the mean for each condition.

4.5.1 Method

Participants

Forty-six undergraduates from the University of California, Berkeley participated in the experiment in exchange for course credit.

Stimuli

The stimuli used in this experiment were the same as those used in Experiment 1.

Procedure

The basic procedure was the same as Experiment 1, except in the reproduction phase. Each participant's responses in this phase formed three chains of 40 trials. The chains started with three original stimuli with width values of 2.63cm, 4.19cm, and 5.76cm, then in the following trials, the stimuli participants saw were their own reproductions in the previous trials in the same chain. To prevent participants from realizing this fact, chain order was randomized and the Markov chain trials were intermixed with 40 control trials in which widths were drawn from the prior distribution.

4.5.2 Results and discussion

Participants' data were excluded based on the same criteria as used in Experiment 1, with a lower testing score of 70% of optimal performance and one additional criterion relevant to the within-subjects case: participants were also excluded if the three chains did not converge, with the criterion for convergence being that the lower and upper chains must cross the middle chain. After

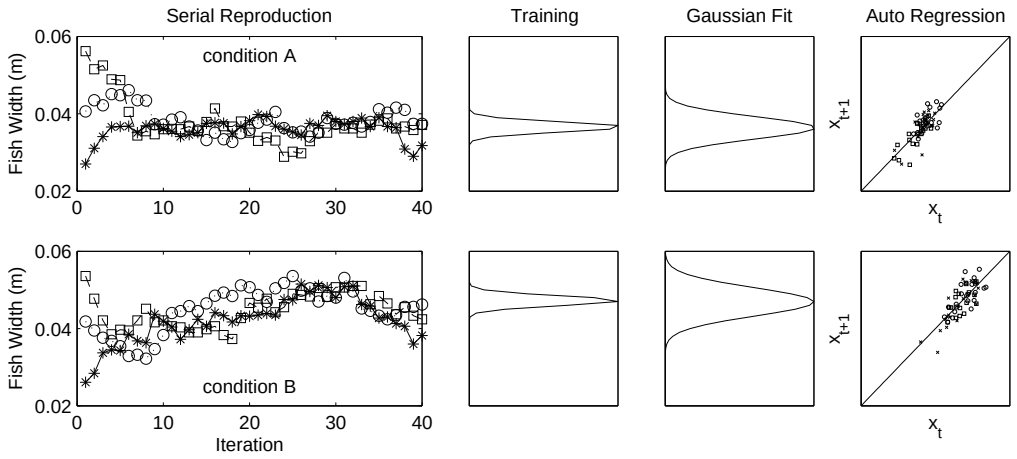

Figure 4.5: Chains and stationary distributions for individual participants from the two conditions in Experiment 2. (a) The three Markov chains generated by each participant, starting from three different values. (b) Training distributions for each condition. (c) Gaussian fits for the last 20 iterations of each participant's data. (d) Autoregression for the last 20 iterations of each participant's data.

these screening procedures, data from 40 of the 46 participants were included, with 21 in condition A and 19 in condition B. It took most participants about 20 trials for the chains to converge, so only the second half of the chains (trials 21-40) were analyzed further.

The locations of the stationary distributions were measured by computing the means of the reproduced fish widths for each participant. For conditions A (3.66cm) and B (4.72cm), the average of these means was 3.32 and 4.01cm respectively ($t(38) = 2.41$, $p = 0.021$). The right panel of Figure 4.4 shows the mean values for these two conditions. The basic prediction of the model was borne out: participants converged to distributions that differed significantly in their means when they were exposed to data suggesting a different prior. However, the means were in general lower than those of the prior. This effect was less prominent in the control trials, which produced means of 3.63 and 4.53cm respectively.

Figure 4.5 shows the chains, training distributions, the Gaussian fits and the autoregression plots for the second half of the Markov chains for two participants in the two conditions. Figure 4.6 shows the autoregression and bias plots for all participants. As in Experiment 1, the autoregression plots show a strong linear relationship between the stimulus x_n and the reconstruction x_{n+1}. Again, the correlation was high in both conditions, with mean values being 0.90 and 0.81 for conditions A and B respectively. The correlations are significant ($p < 0.001$) for all participants except for one in each condition, indicating that the AR(1) model's predictions are highly correlated with the data generated by each participant. The mean partial correlation between x_{n+1} and x_{n-1} given x_n was low, being 0.07 and 0.11 for conditions A and B respectively, suggesting that the Markov assumption was satisfied. The partial correlation was significant ($p < 0.05$) for only one participant

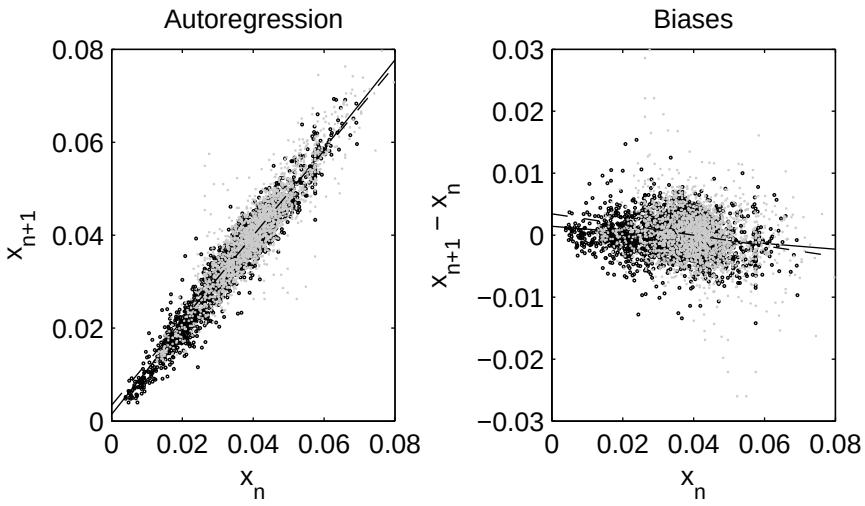

Figure 4.6: Autoregression and bias plots for the two conditions in Experiment 2. (a) The autoregression plots for conditions A (black) and B (gray). (b) The bias plots for conditions A (black) and B (gray).

in condition B. Similar to Experiment 1, the effect of bias was shown in significant ($p < 0.05$) negative slopes in the bias plot, that is, negative correlations between x_n and $(x_{n+1} - x_n)$. This was true for all participants except for two in condition A and one in condition B.

The results of this experiment corresponded well with those of Experiment 1, showing that serial reproduction has similar consequences whether it takes place between-subjects or within-subjects. This correspondence is consistent with our analysis, in which the only factor that is relevant to the convergence of serial reproduction is the repeated reconstruction of stimuli, regardless of whether those reconstructions come from multiple people or just one person. These results suggest that serial reproduction can be used not just for exploring cultural biases, but for investigating the memory biases of individuals. While we examined only a simple form of bias in this experiment – the bias towards the mean of a trained category – we should expect that memory biases that appear small on a single trial can be magnified through the process of serial reproduction, providing a useful lens for exploring the general nature of these biases as well as individual differences.

The correspondence between the results of Experiments 1 and 2 shows up not just in the way in which human behavior conforms to our model predictions, but also in the way in which it deviates. While reconstructions tended towards distributions affected by the prior established through training in both experiments, reflected in the difference in the means of the stationary distributions produced in the two conditions, the actual means of the stationary distributions differed systematically from the means of the distributions used in training. Both experiments produced stationary distributions with means lower than those of the training distributions. This phenomenon suggests that there could be another source of prior other than the ones participants were trained on, a

possibility that we explore in Experiment 3.

4.6 Experiment 3: Serial reproduction with no training on priors

Experiments 1 and 2 tried to establish different prior distributions for reconstruction from memory by providing training on different categories of fish. One possible explanation for why the serial reproduction chains did not converge to stationary distributions that matched these training distributions is that the training might have been insufficient to establish strong prior beliefs. In particular, people may have combined existing expectations about the width of fish with the information provided by the training data when they formed their estimates of the distributions associated with the different categories.

Hemmer and Steyvers (2008) explored how prior knowledge influences reconstruction from memory, and provided a detailed exploration of how such priors might be defined at multiple levels, such as category and object levels. They conducted experiments in which participants were asked to reconstruct the size of familiar objects such as fruits and vegetables. The results showed that reconstructions of objects (e.g., apples) were biased towards the mean size of the superordinate category (e.g., fruit), and at the same time, the size of a smaller version of an object was overestimated at reconstruction while the size of a larger version of the same object was underestimated, reflecting biases at the levels of objects themselves. To explain this effect, they extended Huttenlocher et al.'s (2000) model and proposed a hierarchical Bayesian account of these effects, defining priors at multiple levels.

This analysis suggests a simple account of the systematic differences between training distributions and the stationary distribution of serial reproduction seen in our experiments: people may already have a prior for the width of fish defined at the superordinate level, and are guided by these expectations when learning about fish-farm fish in our experiments. We can explore this possibility by using the serial reproduction method to investigate the priors that people use when they receive no training. We thus conducted an experiment in which participants performed exactly the same serial reproduction task as in Experiment 2, but they were not given the training phase. Since there is no training, our Bayesian model predicts that this will converge to a distribution that reflects people's general knowledge about the width of fish. If the biases seen in Experiments 1 and 2 are a result of these superordinate-level expectations, we should expect the mean of this distribution to be smaller than that of the training distributions used in the two experiments, allowing it to exert an additional influence on reconstructions that leads to a stationary distribution with a lower mean even when training is provided.

4.6.1 Method

Participants

Thirty-four undergraduates from the University of California, Berkeley, participated in the experiment in exchange for course credit.

Stimuli

The stimuli used in this experiment were the same as those used in Experiments 1 and 2.

Procedure

The basic procedure was the same as Experiment 2, except that there were no training and testing phases.

4.6.2 Results and discussion

Participants' data were excluded based on the second, third, and fourth criteria used in Experiment 2. There were no testing scores since no training was given. This resulted in the data from 28 of the 34 participants being subjected to further analyses. As in Experiment 2, only the final 20 responses produced by each individual were analyzed, as it took approximately 20 trials for the chains to converge, and only those data with chains converged were analyzed. The mean width of the fish produced in these trials was 3.54cm, significantly less than the mean of the initial values of each chain, 4.19cm ($t(27) = 4.33$, $p < 0.001$). To preclude the possible explanation of these lower values where the chains end up as simply a downward bias, we also analyzed the difference scores of starting and ending values of the upper and lower chains, computed as the mean of the last 20 trials. The mean values of these two difference scores were 2.21cm and -0.91cm for the upper and lower chains, respectively. All the chains with the lower starting point (2.63cm) moved to higher values, except for two participants. The two-sample t-test also showed significant difference between these scores ($t(54) = 14.62$, $p < 0.001$). These results indicate that people seem to have an *a priori* expectation that fish will have widths smaller than those used as our category means, suggesting that the deviations from the training distributions observed in Experiments 1 and 2 are the consequence of using a prior that is a compromise between this superordinate-level expectation about the width of fish and the training data.

4.7 Experiment 4: Serial reproduction with bimodal distributions

In Experiments 1 and 2, we trained people on simple Gaussian distributions to show that serial reproduction converges to a distribution that reflects memory biases consistent with Gaussian pri-

ors. However, our original analysis of serial reproduction made no assumptions about the form of the priors, indicating that convergence to the prior should be expected in all cases under our assumptions about the process of storage and reconstruction. To test if we observe similar results with people for priors beyond simple Gaussian distributions, we conducted another experiment in which participants were trained on categories with bimodal distributions, that is, they were given bimodal priors. This experiment allows us to determine whether serial reproduction just converges to a Gaussian distribution independent of the prior – a reasonable alternative hypothesis – or is sensitive to the form of the prior distribution in a way that produces bimodal stationary distributions.

Since learning a single category distribution that is multimodal could be challenging (e.g., McKinley & Nosofsky, 1995), we trained people on two unimodal distributions corresponding to the width of fish of different species, where species was unambiguously indicated through the color of the fish (red or blue). The width of fish in each species followed a Gaussian distribution, and these distributions were selected so that the overall distribution of widths was bimodal. We then asked people to produce reconstructions of fish that were glimpsed briefly in "darkness," where there was insufficient light to see the color of the fish. Under these circumstances, reconstructions should be made using the overall distribution of widths, providing a bimodal prior. We called this the *no color* condition.

Examining serial reproduction for bimodal priors also allows us to investigate an issue that has arisen in work following up on Huttenlocher et al.'s (2000) original Bayesian analysis of reconstruction from memory. Sailor and Antoine (2005) conducted experiments of reconstruction from memory in which people were simultaneously trained on two distinct category distributions, and found that even when people were reconstructing two distinct categories they tended to produce reproductions biased toward the overall mean of the stimuli rather than the means of the individual categories. They suggested that this showed an effect of experimental context, which determines the overall range of the stimuli, rather than an effect of categories on reproduction.

We explored whether people produce reconstructions based on experimental context or on category information by adding a second condition to our experiment, in which people reconstructed the width of the fish as before, but now the colors of the fish were visible. We called this the *color* condition. If people use a prior appropriate to the category indicated by the color of the fish, we should expect the serial reproduction chains for fish of different colors to converge to different distributions, reflecting the training distributions for those categories. The overall stationary distribution should also be equivalent to the stationary distribution produced in the *no color* condition. If people simply produce reconstructions using a single distribution derived from the experimental context, we should expect no difference between these chains, since they should converge to the same stationary distribution regardless of category. The stationary distribution should thus be different from that observed in the *no color* condition.

4.7.1 Method

Participants

Eighty-five undergraduates from the University of California, Berkeley, participated in the experiment in exchange for course credit.

Stimuli

The stimuli used in this experiment were the same as those used in the previous three experiments, except that the training distributions were different. The widths of the two types of fish (red and blue) were normally distributed with $\mu_1 = 3.66$cm, and $\mu_2 = 4.72$cm, and $\sigma_1 = \sigma_2 = 0.13$cm.

Procedure

The basic procedure was the same as Experiment 2, consisting of three phases: training, reproduction, and testing.

The training phase was a categorization task as in Experiments 1 and 2. Participants were taught to discriminate two types of fish-farm fish, the red fish and the blue fish. Participants saw fish in "darkness" (grey fish), and guessed the color (blue or red). They then received feedback about the color of the fish. The number of trials and criterion for success was the same as that of the previous experiments.

The reproduction phase consisted of 4 chains of 40 trials, starting from 2.63 and 5.75cm. In the *no color* condition, all the fish were shown in grey. In the *color* condition, two chains were presented in red and another two other chains were in blue. As with the previous experiments, the chains were randomized and mixed with 40 control trials.

The test phase was the same as Experiments 1 and 2, in which participants judge the category of fish stimuli as in the training phase, but no feedback was given.

4.7.2 Results and discussion

Participants were excluded based on the same criteria used in Experiment 2. The data from 70 of the 85 participants passed these criteria and were analyzed further, with 35 in each condition (*no color* and *color*). Again, only the second half of each chain was analyzed.

Figure 4.7 shows the chains produced by two participants in the two conditions, as well as the Gaussian fits of the training data and last 20 trials of the reproduction chains. To test whether people had converged to unimodal or bimodal distributions, we fit each individual's data using a single Gaussian and a mixture of Gaussians with two components. Fitting was done by maximum-likelihood estimation. For the *no color* condition, the parameters of the mixture of Gaussians were estimated using the Expectation-Maximization algorithm (Dempster, Laird, & Rubin, 1977), since the assignment of individual data points to Gaussian components was a latent variable. For the

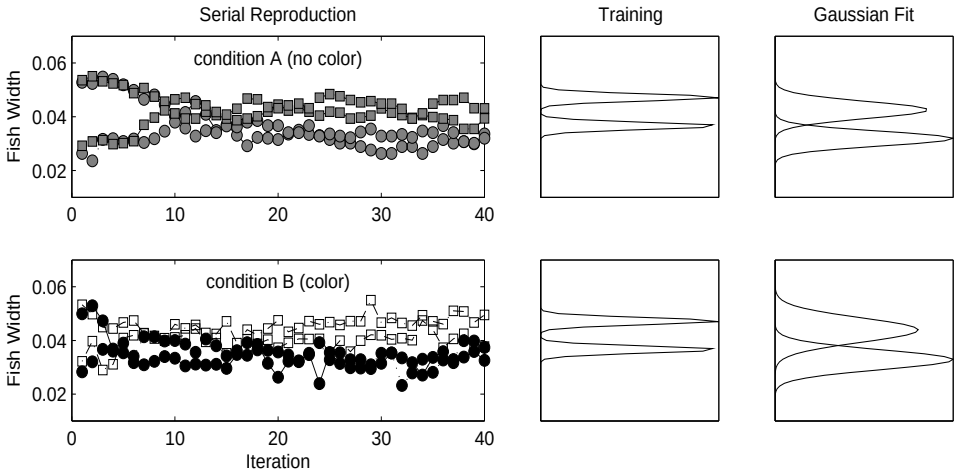

Figure 4.7: Chains and stationary distributions for individual participants from the two conditions in Experiment 4. (a) The four Markov chains generated by each participant, starting from four different values. In condition A, black and white plot markers are the red and blue fish chains respectively, and in condition B, all the chains are gray fish. (b) Training distributions for each condition, black and light-gray plots being the red and blue fish respectively. (c) Gaussian fits for the last 20 iterations of each participant's data. Black and light gray plots in condition A are the red and blue fish chains, and both plots in condition B are the gray chains.

color condition, separate Gaussians were estimated from the responses in the red and blue chains independently, with the result being a mixture of the two distributions.

A better fit by the mixture of Gaussians than the single Gaussian would provide evidence in favor of the prediction that serial reproduction converges to a distribution reflecting the bimodal prior in the *no color* condition, and that it converges to different distributions for different categories (rather than a single distribution guided by context) in the *color* condition. We compared the fit of the two models in each condition using likelihood-ratio tests. Since the single Gaussian model is a special case of the mixture of Gaussians, twice the difference between the log-likelihoods of the two models should follow the χ^2 distribution with degrees of freedom equal to the difference in the number of parameters of the models under the assumption the simpler single Gaussian model is true (Rice, 1995). To confirm the results of the likelihood-ratio tests, we also computed the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) values as model selection measures.

In the *color* condition, likelihood-ratio tests showed that the data of 34 out of 35 participants were significantly better fit by the two Gaussian model (all $p < 0.05$). Both AIC and BIC values indicated exactly the same result. The average values of the means of the two Gaussians for those 34 participants are 3.10 and 3.91cm, and a two-samples t-test showed that these means are significantly different ($t(66) = 5.48,\ p < 0.001$). These results indicate that for the majority of participants in the "color" condition, where serial reproduction was done with the context of the

color of each category present, the chains converged to bimodal distributions reflecting the priors established through training.

In the *no color* condition, one participant's data produced degenerate results for maximum-likelihood estimation and were omitted from further analysis. Likelihood-ratio tests showed that the data of 18 out of the remaining 34 participants were significantly better fit by the two Gaussian model (all $p < 0.05$). The average values of the means of the two Gaussian components for those 18 participants' data were 3.00 and 4.08cm, and a two-samples t-test showed that these means are significantly different ($t(34) = 4.67$, $p < 0.001$). Thus for about two thirds of the participants serial reproduction converged to a bimodal distribution. The AIC and BIC values showed converging results: 22 and 15 out 34 participants' data were better fit by the two Gaussian model using the AIC and BIC criteria respectively, consistent with the greater conservatism that the BIC displays towards more complex models.

These results support two conclusions. First, the finding that at least some of our participants produced bimodal stationary distributions in the *no color* condition indicates that serial reproduction is sensitive to the form of the prior and not just its mean and variance. This complements our observation of convergence to the prior in previous experiments, showing that this property of serial reproduction generalizes beyond simple Gaussian priors. We do not view the fact that between a half and two thirds of participants (depending on model selection measure) produced bimodal distributions as a major problem, since our primary goal was to provide an existence proof for convergence to other distributions, and a variety of factors including overestimating the variance of the categories and the relatively small number of trials in each chain could have resulted in the stationary distributions appearing unimodal. Second, the overwhelming tendency for people to converge to stationary distributions best characterized by two Gaussians in the *color* condition indicates that their reconstructions were guided by the category of the stimulus rather than general experimental context, contrary to the claims of Sailor and Antoine (2005). This conclusion is further supported by the production of bimodal stationary distributions even in the *no color* condition, where the absence of category cues would presumably strengthen the reliance on general experimental context.

4.8 General discussion

We have presented a Bayesian analysis of serial reproduction, providing both general predictions about the outcome of this process for arbitrary stimuli and specific predictions for the special-case of one-dimensional stimuli with Gaussian priors. The results of our four experiments confirm the predictions produced by this analysis. Experiments 1 and 2 showed, in both within-subject and between-subject cases, that serial reproduction using one-dimensional stimuli with Gaussian priors converged to a distribution consistent with the prior on which participants were trained. Using serial reproduction without training, Experiment 3 revealed people's general expectations about the size of fish were consistent with a systematic bias observed in Experiments 1 and 2. In Experiment 4, we tested whether the predictions of our Bayesian analysis held beyond simple

Gaussian distributions by establishing bimodal distributions as priors. The results confirmed that serial reproduction converged to bimodal distributions reflecting those priors, and also tested the hypothesis that reconstruction is biased by the context of the stimuli. The results of this experiment showed that category structure and not general experimental context seemed to be guiding people's reconstructions. Those results also demonstrate an effective and more sensitive way than the tranditional methods of tapping people's prior knowledge of the world.

The theoretical analysis and experimental results we have presented in this paper make contact with an experimental literature on reconstruction from memory and memory biases revealed through serial reproduction that goes back to the 1930s. However, this phenomenon has not been formally analyzed before and most previous research used complex stimuli that are hard to control and difficult to interpret. To our knowledge, ours is the first detailed mathematical analysis of serial reproduction, and the first formal confirmation of Bartlett's (1932) conclusion that the outcome of serial reproduction reflects people's biases. Our use of simple one-dimensional stimuli allowed us to develop an even more precise model based on first-order autoregressive processes, and provided way to develop a well-controlled experimental method that could be used in a quantitative test of the predictions of our model. However, the Markov chain analysis also generalizes to any kind of stimuli, hypotheses, and prior distribution, opening a lot of opportunities for further exploration of the relationship between memory biases and serial reproduction. In the remainder of the paper we highlight some connections to other research and consider the limitations and possible future directions of this work.

4.8.1 Connections to other research

The work we have presented here has connections to two other lines of research exploring how ideas from Bayesian statistics can be used to understand human cognition: models of reconstruction from memory, and iterated learning. We will briefly summarize these two sets of connections in turn.

Reconstruction from memory

As discussed above, previous papers have proposed a Bayesian analysis of reconstruction from memory. Huttenlocher et al. (2000) proposed that reconstructions should be a compromise between the observed value and the mean of a category in order to minimize reconstruction error. The resulting model is equivalent to that obtained by treating the problem as one of Bayesian inference with a Gaussian prior. Hemmer and Steyvers (2008) took an explicitly Bayesian perspective on this problem and extended Huttenlocher et al.'s analysis to other priors, such as hierarchical priors defined at the level of both individual objects and categories. They explored the memory biases shown with real categories using naturalistic stimuli, showing that people exhibit relatively strong biases in a memory task using these categories. Stewart et al. (2005) explored a similar model in the context of sequential effects on memory recall.

Our primary theoretical contribution in this paper is an analysis of the predictions that this Bayesian account of reconstruction from memory makes about serial reproduction. This analysis extends the scope of the phenomena to which Bayesian models of reconstruction have been applied, but is otherwise consistent with the work of Huttenlocher et al. (2000), Hemmer and Steyvers (2008), and Stewart et al. (2005). The fact that this account produces predictions that are consistent with the conclusions of Bartlett (1932) and with our own experiments when applied to serial reproduction provides further support for its utility as a model of reconstruction from memory.

Iterated learning

The key step in proving that serial reproduction converges to distribution determined by the prior was noting that alternating between sampling μ from the posterior distribution $p(\mu|x)$ and x from the likelihood $p(x|\mu)$ defined a Markov chain with stationary distribution $p(x,\mu) = p(x|\mu)p(\mu)$. While this is a Markov chain of a type commonly used in Bayesian statistics, Griffiths and Kalish (2007) observed that such a process could provide a natural model of the cultural transmission of information. In particular, they showed that the process of *iterated learning*, in which a sequence of people each learns from data generated by the previous person and then generates the data provided to the next person, could be analyzed as a Markov chain of exactly this kind.

Griffiths and Kalish (2005, 2007) analyzed iterated learning for a sequence of Bayesian learners, each of whom forms a hypothesis based on the data generated by the previous learner. If learners sample hypotheses h from the posterior distribution $p(h|d)$ and data d from the likelihood $p(d|h)$, the result is a Markov chain that converges to $p(d,h) = p(d|h)p(h)$. As a consequence, the probability that a learner selects a particular hypothesis h on a given iteration converges to the prior probability of that hypothesis, $p(h)$.

While iterated learning was originally proposed as a way to model language evolution (Kirby, 2001), the prediction of convergence to the prior is interesting in the context of cultural evolution more generally, since learning is one of the ways in which information is transmitted between people. It also suggests that we might be able to identify the biases that guide human learning by reproducing the process of iterated learning in the laboratory. This basic prediction has been confirmed through experiments with human participants showing that iterated learning of functions (Kalish, Griffiths, & Lewandowsky, 2007) and categories (Griffiths, Christian, & Kalish, 2008) results in an increase in the prevalence of concepts that are easy to learn (ie. those that have high prior probability).

Iterated learning and serial reproduction have a basic structural correspondence, both being concerned with the transmission of information along a chain of individuals. Both are thus instances of a paradigm known as a "diffusion chain" in the broader anthropological literature (for a review, see Mesoudi, 2007). The key difference between the two is the mechanism of transmission – the kind of cognitive process involved. In iterated learning this mechanism is learning, while in serial reproduction it is memory. The connection between the two paradigms that we draw on here results from treating both learning and memory as inductive problems that can be solved via

Bayesian inference.

The results we present in this paper thus complement the work on iterated learning mentioned above, showing how similar theoretical analyses and empirical findings hold for transmission of information via reconstruction from memory. This broadens the scope of cultural transmission phenomena we might hope to explain, as well as the range of psychological biases we have the opportunity to investigate. It also provides a link to an empirical literature within psychology that goes back over 70 years, and a way to validate Bartlett's (1932) original claims about the effects of serial reproduction.

4.8.2 Limitations and future directions

While our theoretical results apply for arbitrary stimuli and prior distributions, the experiments we presented in this paper used only one-dimensional stimuli and prior distributions that can be expressed as mixtures of Gaussians. Our choice to use these stimuli and priors was motivated by a desire for simple, well-controlled experiments about which we could make clear quantitative predictions and the existing empirical literature based on similar assumptions (Huttenlocher et al., 2000; Hemmer & Steyvers, 2008; Stewart et al., 2005). However, an important direction for future research will be examining how well our theoretical results hold for other stimuli and kinds of memory biases.

One reason to explore serial reproduction with other stimuli is to provide further test of the predictions produced by our Bayesian analysis. To do so, we would ideally use stimuli for which memory biases are already well established. For example, Feldman (2000) used a reconstruction task to investigate biases for boolean concepts, building on the work of Shepard, Hovland, and Jenkins (1961). In this task, people were shown a division of a set of objects varying along binary dimensions into two groups, and then asked to reconstruct the division from memory. The ease of reconstruction varied with the complexity of the rule that described the division. We should expect the same memory biases to manifest in serial reproduction, with the lower-complexity rules being more likely to survive the process. Understanding how concepts change when passed from one person to another is particularly interesting in light of recent work exploring the stability of different kinds of religious concepts under cultural transmission (Barrett & Nyhof, 2001; Boyer & Ramble, 2001).

Another reason to conduct experiments with other stimuli is that our results justify using serial reproduction as a method for investigating memory biases. Since these biases have an effect each time people reconstruct a stimulus from memory, Serial reproduction can magnify what might be small effects in the context of a standard memory task. Controlled experiments in serial reproduction might thus be a valuable tool for exploring memory biases in a variety of domains. While this method has been used heuristically in the past, our results provide it with a more rigorous justification, as well as examples showing that the method works in both between- and within-subjects designs.

4.8.3 Conclusion

We have presented a Bayesian account of serial reproduction, and tested the basic predictions of this account using four controlled laboratory experiments. The results of these experiments are consistent with the predictions of our account, with serial reproduction converging to a distribution that is influenced by the prior distribution established through training. Our analysis connects the biases revealed by serial reproduction with the more general Bayesian strategy of combining prior knowledge with noisy data to achieve higher accuracy. It also shows that serial reproduction can be analyzed using Markov chains and first-order autoregressive models, providing the opportunity to draw on a rich body of work on the dynamics and asymptotic behavior of such processes. These connections allow us to provide a formal justification for the idea that serial reproduction changes the information being transmitted in a way that reflects the biases of the people transmitting it, establishing that this result holds under several different characterizations of the processes involved in storage and reconstruction from memory.

Chapter 5

Reproducing Color Term Universals

5.1 Introduction

Linguistic universals – properties that seem to hold across all human languages – have the potential to provide a unique insight into the nature of human cognition. Universals in systems of color terms are among the best documented of these properties. Languages with a small number of terms show characteristic patterns in which colors those terms label (Berlin & Kay, 1969; Kay & McDaniel, 1978; Kay & Maffi, 1999). For example, when a language has three color terms, one term usually denotes light colors, another dark colors, and the third reddish colors (Kay & McDaniel, 1978; Kay & Maffi, 1999). The existence of these patterns raises an important question: Where do these universals come from? A variety of explanations have been presented, including appeals to perceptual similarity (Davidoff, Davies, & Roberson, 1999; Roberson, Davies, & Davidoff, 2000), the constraints induced by communicating with other people about colors (Steels & Belpaeme, 2005; Baronchelli, Gong, Puglisi, & Loreto, 2010), and the idea that universals reflect optimal partitions of the space of colors based on their physical or perceptual properties (Shepard, 1992; Yendrikhovskij, 2001; Jameson & D'Andrade, 1997; Regier, Kay, & Khetarpal, 2007). We show that patterns similar to those seen across human languages can emerge in systems of color terms merely as a result of transmission from person to person. This suggests that human perceptual and learning biases, brought out through the process of cultural transmission, are sufficient to account for these linguistic universals.

Computer simulations of cultural evolution have been used as one of the main tools for evaluating different accounts of the origins of linguistic universals. If we view language as a system culturally transmitted from generation to generation, this process of transmission provides an opportunity for universals to arise from biases that cause learners to prefer some language structures over others. This hypothesis has been explored using a simple model of cultural transmission known as "iterated learning", in which a sequence of computer-simulated agents each learns from the behavior of the previous agent in the sequence (Kirby, 2001). This establishes a situation similar to the children's game of "Telephone", where the information being transmitted is altered as

it is passed from one agent to another. Mathematical analyses of iterated learning show that over time, the information being transmitted gradually changes to become consistent with the biases of the agents involved (Griffiths & Kalish, 2007; Kirby, Dowman, & Griffiths, 2007) (see Section 5.2 for details). These results suggest that if systems of color terms similar to those seen in human languages emerge from a process of cultural transmission by iterated learning, the biases of individual learners may be sufficient to account for the observed linguistic universals.

To apply the iterated learning model to the transmission of systems of color terms, we need to construct a sequence of agents in which each agent learns a system of color terms from examples provided by the previous agent, and then generates examples which are provided to the next agent in the sequence. Previous work has used computer simulations to demonstrate that iterated learning with computer-simulated agents can produce systems of color terms that mirror universal patterns seen in human languages (Dowman, 2007, 2009). However, the conclusions drawn from these simulations are limited by the extent to which the agents involved accurately mimic human perceptual and learning biases. We addressed this problem by conducting a large-scale laboratory experiment based on iterated learning. In our experiment, human learners acquire and transmit novel systems of color terms, creating a laboratory simulation of the process of cultural evolution. We examined how these systems of color terms change over time, comparing the results to those seen in human languages. We found that the systems of color terms generated by our iterated learning chains converged over time to become similar to those seen in non-industrial human societies.

The most definitive account of the regularities seen in systems of color terms is provided by the World Color Survey (WCS) (Kay et al., 1991, 1997, 2009). In the WCS a total of 330 color chips, comprised of 40 equally spaced Munsell hues at 8 levels of lightness and achromatic chips at 10 levels of lightness (see Figure 5.1 (a)), were presented to speakers of 110 different languages, all from non-industrial societies. Those speakers were asked to name each color chip, and also to point out the most representative chip for each color term. We thus used stimuli based on those from the WCS, with the goal of comparing the languages produced by our laboratory participants with those of the speakers of the WCS languages (see Section 5.3 for details). In order to limit the influence of the English color terms known by our experimental participants, we made it clear to participants that they were learning a novel artificial language. We also fixed the number of terms in the artificial languages to between two and six terms, consistent with the number of terms seen in the majority of WCS languages, but significantly different from the eleven basic color terms of English. This forced the participants to explore unfamiliar ways of partitioning color space.

Following the iterated learning model, we simulated the cultural transmission of languages by using the responses of one participant to train the next (see Section 5.3 for details). Each participant was trained by being shown a set of randomly selected example colors, each paired with the word chosen to name it by the previous participant. The first participant in each chain was shown examples sampled from a computer-generated initial partition. After training, participants were asked to label all 330 WCS colors based on the examples they had seen. A total of 20 chains were run using randomly generated initial partitions of the 330 colors, corresponding to four chains for each of two, three, four, five, and six terms. Use of a random initial partition allowed us to examine what structure people added to the system. In order to determine whether the chains had

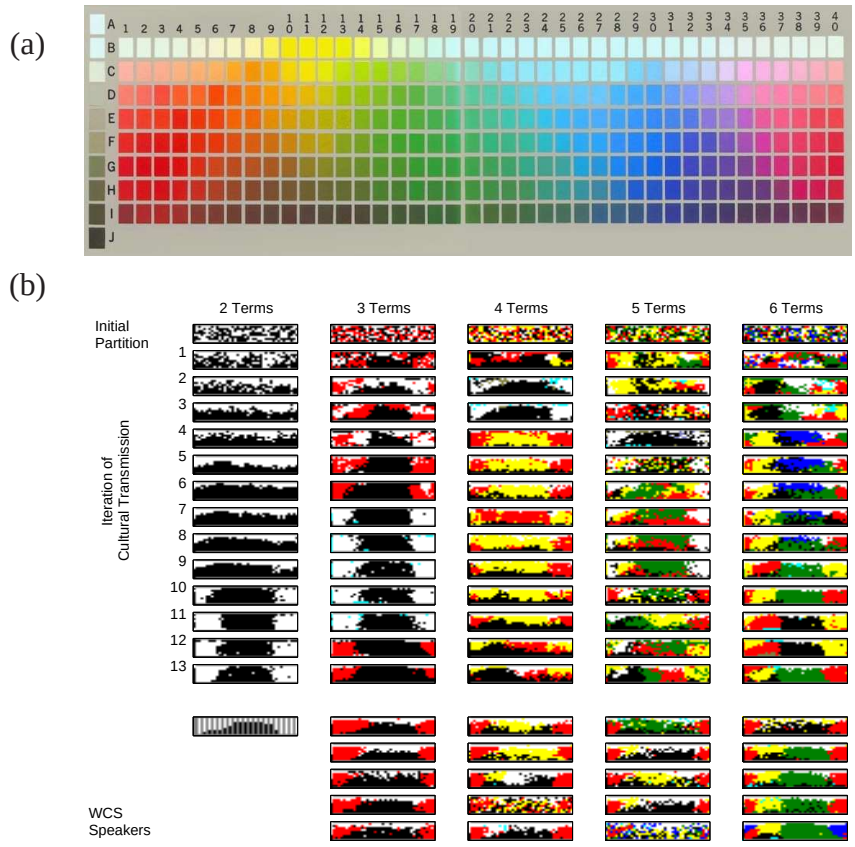

Figure 5.1: Simulating the cultural transmission of color term systems. (a) The array of colors used in the World Color Survey (WCS). Members of 110 non-industrial societies reported the terms that their languages used to label these colors. (b) Examples of color-term systems produced by simulating cultural transmission in the laboratory. Each column shows one chain of systems produced by participants learning novel color terms from examples sampled from the system of labels assigned to colors by the previous participant. Transmission proceeds down the column, and different columns show chains for systems with two, three, four, five, and six terms. The color chips are arranged in the same order as in (a), so the position of a chip in the arrays corresponds to its color. The colors signify which chips were labeled with the same color word, and do not directly correspond to the colors denoted by the words (light blue is used to indicate a minority term, used for fewer than 5% of all chips). The first system in each column is a randomly generated initial partition. The last system in the first column is the Dani language, for which only aggregate data using a subset of the WCS array are available (the gray bars correspond to unlabelled chips). The last five systems in the remaining columns are data from individual speakers of the closest matching WCS language (determined by averaging Variation of Information (VI) values across iterations 4-13 of each chain).

converged, we also ran ten chains initialized with partitions that were more characteristic of human languages (see Section 5.5.2 for additional analyses using these chains). Each chain included 13 participants.

5.2 Background on iterated learning

We can analyze the process of iterated learning by assuming that our learners are rational Bayesian agents. In this framework, learners come up with the *posterior* probability $P(h|d)$ of a hypothesis h given the observed data d by applying Bayes' rule,

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')} \tag{5.1}$$

where $P(d|h)$ is the *likelihood*, indicating the probability of observing d if h were true, and $P(h)$ is the *prior* probability, indicating the extent to which the learner was willing to accept h prior to observing d. The prior encodes the learner's inductive biases, and is a factor that combines with the observed data to yield a conclusion.

In iterated learning, data are passed along a chain of learners, and we assume that the same Bayesian inference process happens repeatedly at each generation. Each learner samples a hypothesis from their posterior distribution, and then generates data by sampling from the likelihood function associated with that hypothesis. This can be analyzed as a Markov process: The probability with which each learner selects a particular hypothesis is independent of the data produced by all generations except the immediately preceding one. Griffiths and Kalish (Griffiths & Kalish, 2007) showed that when learners share a common prior distribution, as the process of iterated learning continues, the probability with which a learner selects a hypothesis converges to the prior probability of that hypothesis as the process of iterated learning continues. Likewise, the probability of generating data d converges to the prior predictive distribution, being the average of the likelihood over the prior, $P(d) = \sum_h P(d|h)P(h)$. Similar effects are observed with other Bayesian models of learning (Kirby et al., 2007).

The convergence of iterated learning to the prior potentially provides an explanation of linguistic universals, including universals in color naming. Languages are constantly being passed from speaker to speaker via a process of cultural transmission similar to iterated learning, so if this process provides a way for perceptual and learning biases of the kind captured by a prior distribution to have an effect on the structure of languages, we should expect human languages to have properties that mirror these human biases. If this hypothesis is correct, we should expect to see systems of color terms transmitted via a process of iterated learning to change over time to resemble those that appear in the WCS.

5.3 Materials and methods

5.3.1 Participants

Participants were 399 members of the community at the University of California, Berkeley, receiving either course credit or approximately $10/hr for taking part in the experiment. Participants had normal color vision.

5.3.2 Stimuli

Each participant learned a system of color terms from examples of colors and the terms that were associated with them, and they then generalized those terms to new colors. A total of 330 colors were used as stimuli, corresponding to the computer screen analogues of the 330 Munsell color chips used in the WCS. Each term was a randomly-allocated pseudo word consistent with the phonological system of English, and the words were varied randomly across participants. The color stimuli were presented on an Apple iMac computer by a Java program.

Munsell values of the 330 color chips were converted to values in CIE 1931 XYZ space and RGB space using the GretagMacbeth Munsell Conversion Database Version 6.2. We then used those RGB values to present the 330 color chips on the computer monitor. The monitor was calibrated using a ColorVision Spyder2 colorimeter/color calibrator on a regular basis. The accuracy of calibration was assessed using a photometer (Minolta Chroma Meter CS-100, manufactured by Minolta Camera Co., Ltd., Osaka, Japan.), confirming that the range of variation in the stimuli was small enough that no two colors from the stimulus set were confusable.

5.3.3 Procedure

We simulated a total of 30 iterated learning chains, each with 13 “generations” of learners. Each chain varied in the number of terms that were allowed in the “language” being transmitted, with two, three, four, five or six terms per language. The first learner in each chain received data generated from one of three types of initial partition of the WCS color space: hue, lightness, and random. The “hue” and “lightness” partitions were approximately equal vertical and horizontal partitions of the color space into the relevant number of categories. (In the “hue” division, the achromatic chips were grouped with a randomly chosen hue partition.) The “random” partitions were a truly random partition of the color space, with an equal number of instances of each term, and were generated uniquely for each chain. These three kinds of initial partition were used as a means of checking the convergence of iterated learning: By starting the chains with very different systems of color terms, we could easily establish when the influence of the initial partition had disappeared. The following generations of learners all received data generated from the responses of the previous generation, as detailed below. We ran a total of 20 random chains, four for each number of terms, and five hue and five lightness chains, one for each number of terms.

Each participant was trained on the system of color terms by being shown a set of colors together with the corresponding terms. The total number of observed colors was six times the number of terms in the language. These chips were chosen at random from the 330 colors making up the full array, and were labeled according to either the initial partition (for the first learner) or the responses of the previous learner (for subsequent participants). In order to reduce the memory demands of the task, these training examples remained on the screen while the participant went on to label all 330 colors from the WCS array. On every trial, the participant was presented with a color, and was then asked to select one of the terms to label that color. No feedback was given during this phase of the experiment. The responses of each participant thus produced a partition of the set of 330 colors, and this partition was used to generate the labels given to colors for the next learner in the chain.

To test the hypothesis that our iterated-learning data simply converges to English, we also conducted another experiment in which nine English-speaking participants were asked to label the 330 WCS colors with one of the six English color terms (*black*, *white*, *red*, *green*, *yellow*, and *blue*). The experiment was conducted in exactly the same environment as the iterated learning experiment. On each trial, the participant was presented with a color and was asked to select one of the English terms to classify the color.

5.3.4 Comparing partitions

Analyzing the results of our experiment presented a challenge: how could we evaluate whether two systems of color terms were similar? Various methods have been proposed for solving this problem. For example, Kay and Regier (Kay & Regier, 2003) converted the color chips from Munsell space to CIE L*a*b* space so they could compute the centroid for each color term. Centroid distances could then be used to compare clusterings. However, just using centroid measurements might have discarded important information about the variance of a cluster, and about the locations of boundaries. This method is also dependent on the psychological validity of the CIE L*a*b* representation of colors, which is disputable (Dowman, 2007).

Since our participants' responses consisted of partitions of the same set of colors as those used in the WCS, we compared the Munsell arrays directly, without referring to another color space. Our technique used an information- theoretic measure known as *Variation of Information* (VI), which is able to quantify the similarity of alternative clusterings of a set of items (Meilă, 2007). Given two clusterings C and C', the VI is

$$VI(C,C') = H(C) + H(C') - 2I(C,C') \tag{5.2}$$

where $H(C)$ is the *entropy* of C,

$$H(C) = -\sum_{k=1}^{K} P(k)\log P(k) \tag{5.3}$$

where k ranges over the cluster labels and $P(k)$ is the probability of an item being assigned to each cluster, and $I(C,C')$ is the *mutual information* between the two clusterings

$$I(C,C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P'(k')} \tag{5.4}$$

where $P(k,k')$ is the probability an item belongs to cluster k in clustering C and to k' in clustering C'.

5.4 Main results

Figure 5.1 (b) shows one set of chains initialized with random partitions, with the number of terms varying from two to six. Through this simple visualization of the data, we can see that each chain started from an unnatural color-term system, and that transmission along the chains resulted in a rapid restructuring towards a more regular form. We measured the similarity between the systems of color terms produced by our participants and those in the WCS using Variation of Information (VI) (Meilă, 2007), an information-theoretic measure of the distance between two partitions (see Section 5.3 for details). The VI value for two systems of color terms was calculated by comparing the relative frequencies of the terms in the two systems, as well as the extent to which they partitioned the color space in the same way. A high VI value reflects a larger difference between two color term systems, whereas a small VI value indicates that the two systems are more similar. At the bottom of each chain we show five randomly selected speakers from the closest matching system from the WCS according to this measure. Since there was no two-term language in WCS, we compared the two-term case to the Dani language (Heider, 1972). To provide a quantitative analysis of the similarity between the systems of color terms produced by our participants and those observed in the WCS data we calculated the mean distance between these two sets of partitions. We first calculated the VI between each system produced by our participants and each system produced by the speakers of each language in the WCS, and then averaged across all speakers within each WCS language, then across all languages.

Figure 5.2 (a) shows the results of our comparison between the systems of color terms produced by our participants and the WCS. The distance between chains with random initial partitions and the WCS clearly decreases across iterations. A paired t-test on the VI values for the initial and final systems in the 20 random chains showed a statistically significant difference ($t(19) = 5.83, p < 0.001$). The remaining question is how close our data are to the WCS data: What counts as a low VI score? To address this question, we computed the VI between all pairs of languages in the WCS. The average pairwise VI is shown in Figure 5.2 (a). This average is extremely close to the mean VI seen in our random chains once they converge. The difference between the VI scores for the systems produced by the final participants in each of our random chains and the VI scores for individual speakers of languages from the WCS was not statistically significant by a two-samples t-test ($t(128) = 0.09$, $p = 0.93$), and in fact no statistically significant differences were observed

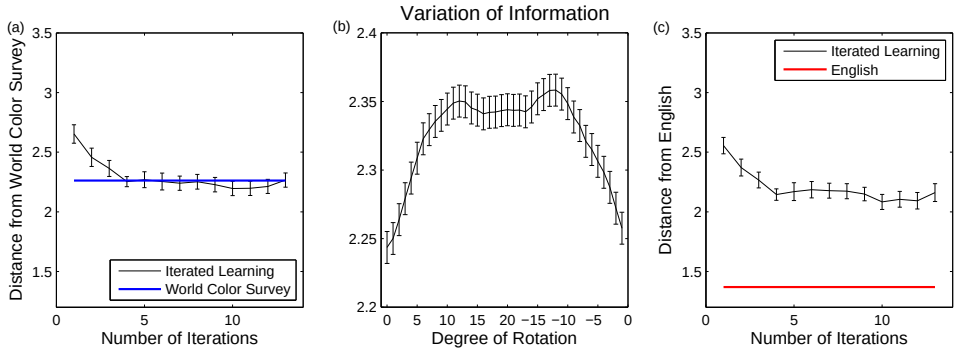

Figure 5.2: Comparing systems of color terms produced by simulating cultural transmission in the laboratory with those from non-industrial societies, as represented by the World Color Survey (WCS) and English color systems. The distance between systems is assessed using Variation of Information (VI), an information-theoretic measure of the difference between two partitions of the same set of elements. (a) The mean VI between systems produced by English-speaking participants and those seen in the WCS decreases as the number of iterations of cultural transmission increases. The VI converges to a level that is not statistically significantly different from the mean VI between languages in the WCS. (b) Taking the final systems produced by our laboratory participants, rotation along the hue dimension of the WCS array results in a worse fit to the WCS, confirming that there are significant similarities between the way the languages emerging in our experiments, and those spoken in non- industrial societies, categorize colors in terms of the hue dimension in the WCS. The horizontal axis indicates the number of steps the chips were rotated to the right along the hue dimension of the array shown in Figure 1. (c) The mean VI between systems produced in the iterated-learning experiment and those produced by participants applying English color terms. The VI between systems in iterated learning chains and English systems did not reduce to the same level as the mean VI between systems produced by English speakers.

from the fourth iteration onwards. These results suggest that the systems of color terms generated in our laboratory experiments are indeed consistent with the data collected from the WCS (see Section 5.5 for further analysis).

One potential objection to the conclusion that our chains converged to a distribution similar to the WCS could be that the reduction in VI may merely be a result of giving the same label to neighboring colors, producing a more coherent classification of the color space. As the systems of color terms in the random chains move towards more regular forms, the VI scores will go down naturally, regardless of whether the actual partition of terms reflects the structure of the WCS or not. To further test the consistency between the results of our experiment and the WCS data, we compared the degree of match of each system to the WCS when it was rotated in the hue dimension (the horizontal dimension for the array shown in Figure 5.1) by varying amounts. If the systems produced by our participants provide a non-trivial match to the WCS, we would expect

that the more a partition was rotated out of position, the worse the resulting match would be. A similar procedure was previously used by (Regier et al., 2007) to provide evidence for universals in color naming. Figure 5.2 (b) shows the mean VI values for the partitions generated by the final participants in each of our random chains, when rotated from 0 to 20 steps in the hue dimension. Paired t-tests on VI values for no-rotation vs. maximum-rotation ($t(19) = -5.95, p < 0.001$), no-rotation vs. quarter- rotation ($t(19) = -3.45, p < 0.01$), and no-rotation vs. three-quarter- rotation ($t(19) = -4.51, p < 0.001$) all showed statistically significant differences, indicating that the partitions produced by our participants fit the WCS data significantly better than the rotated systems. This analysis thus confirmed that the iterated learning chains did converge to systems reflecting the patterns in the assignment of terms to colors of different hues that is evident in the WCS data.

Another concern might be that the systems produced by our participants are converging to the English color system, since all our participants were speakers of English. The question is then how close our iterated-learning chains are to English color systems. To compare our results with English, we conducted another experiment in the same laboratory environment, and asked English participants to label each of the WCS colors with one of six basic English color terms (see Section 5.3 for details). We then computed the VI values between systems produced by the iterated-learning chains with random initial partitions and those in the English experiment in the same way as we did for the WCS data. Figure 5.2 (c) shows the results of this comparison. Although the distance between iterated- learning chains and English systems decreases across iterations, it was still quite far from the average pairwise VI values among English systems: a two-samples t-test showed a statistically significant difference ($t(27) = 6.98, p < 0.001$) between the VI scores produced by the final participants in each chain and those for the English systems (and likewise for all other iterations). This indicates that the classifications produced by different participants using English terms were more similar to each other than they were to the results of our iterated-learning experiment, demonstrating that participants did not simply apply English color categories when classifying colors in the iterated-learning chains.

5.5 Additional analyses

5.5.1 More detailed analysis of random chains.

In Section 5.4 we report a paired samples t-test comparing the initial and final VI scores for the random chains. A more detailed analysis, examining the effect of number of terms as well as iterations, can be obtained by running a two-way ANOVA with the number of terms as a between-subject factor and iteration as a within-subject factor. The main effects of both number of terms and iterations are significant ($F(4,15) = 11.56, p < 0.001$, and $F(12,180) = 11.48, p < 0.001$, respectively), as is the interaction between these two factors ($F(48,180) = 1.84, p < 0.005$). This is consistent with the patterns apparent in the left panel of Figure 5.3, where it can be seen that systems with fewer terms tend to have lower VI scores, and while VI tends to decrease across iterations this trend manifests differently with different numbers of terms. The t-test reported in

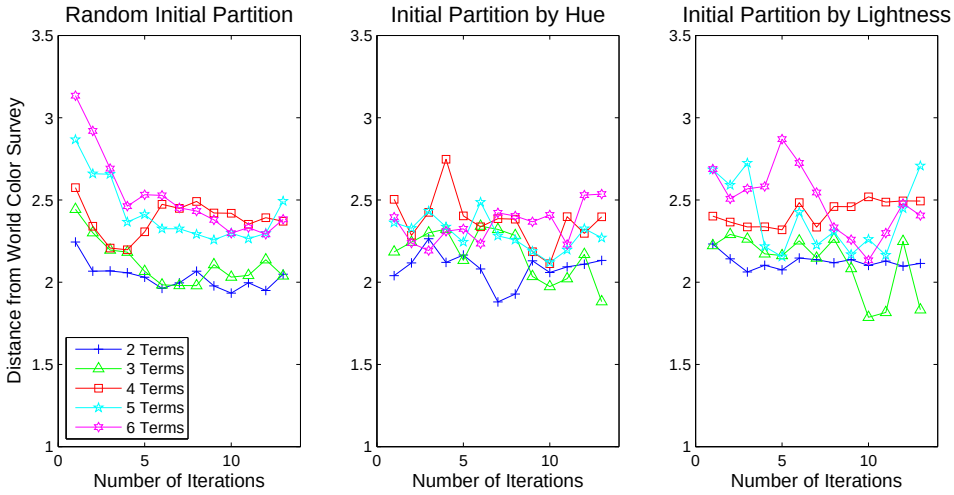

Figure 5.3: Mean distance from the WCS systems for chains with different numbers of terms and different initial partitions.

Section 5.4 shows that the effect of iterations is robust even when the effect of the number of terms is not modeled directly.

5.5.2 Assessing convergence.

In addition to the chains initialized with random partitions reported in Section 5.4, we ran a set of chains initialized with partitions that are closer to human languages (the hue and lightness partitions mentioned in the Section5.3). Our goal in running these chains was to have a way to assess convergence, since we could evaluate when chains from very different starting points began to show similar results. This would also allow us to diagnose if any chains had become stuck at solutions that were still party determined by their initial state. The results for all three types of initial partitions are shown in Figure 5.3. In all cases, chains with the same number of terms produced color term systems with a similar distance from the WCS, suggesting that there were no problems with convergence.

5.5.3 Fixing the number of terms.

The analyses presented in Section 5.4 compared systems of color terms produced by our participants against all languages in the WCS, regardless of the number of terms they used. While there is no problem comparing partitions with different numbers of elements using VI, there may be a concern that languages with different numbers of terms partition color space in different ways, and the distribution of the number of terms in the WCS could introduce artifacts. To address this concern we performed the same analyses using comparisons in which each system was compared only

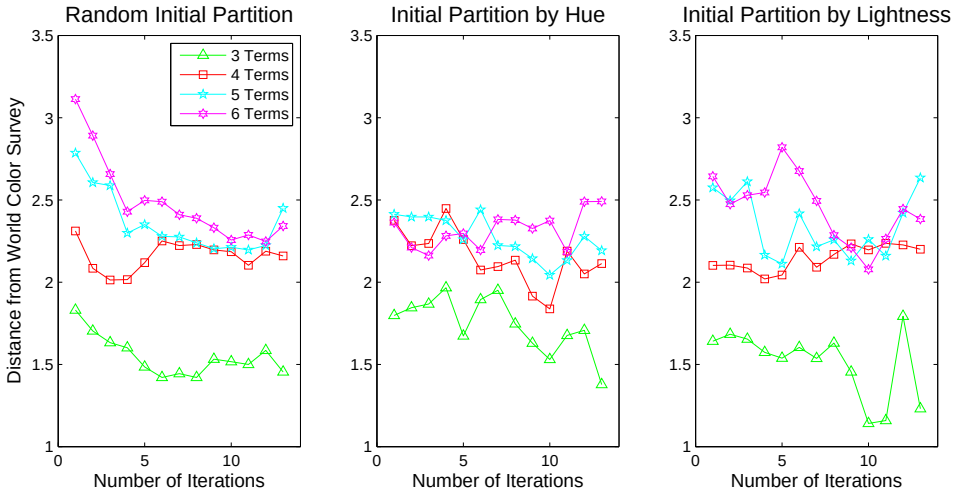

Figure 5.4: Mean VI from the WCS systems for chains with varying numbers of terms and different initial partitions, when comparisons were only made with WCS languages that use an equal number of terms.

against those languages in the WCS with the same number of terms. Because different speakers of the same language sometimes use different numbers of color words (Kay & Maffi, 1999), the number of terms in a WCS language was determined by constructing mode maps (the partition of color space produced by assigning each color its most frequent label across speakers). The number of terms was taken to be the number required to cover 95% of the colors in the WCS array using the mode map for that language. This ensured that we also accommodate noise in the WCS data, and exclude minor terms that were used to label colors that were at the margins of the main color categories. Since no WCS languages used only two terms, we had to omitted the two-term chains from these analyses.

The VI scores as a function of iteration are shown in Figure 5.4. Statistical analyses were consistent with the results given in Section 5.4. A paired t-test on the VI values for the initial and final systems in the random chains showed a statistically significant difference ($t(15) = 5.12, p < 0.001$). The comparison of the VI values for the final systems produced by our participants to the VI values between WCS languages (constraining the latter to only compare languages with the same number of terms, for three to six terms) showed that the difference was not statistically significant ($t(124) = -0.86, p = 0.39$). Paired t-tests on VI values for no-rotation vs. maximum-rotation ($t(15) = -5.14, p < 0.001$), no-rotation vs. quarter-rotation ($t(15) = -3.08, p < 0.01$), and no-rotation vs. three-quarter-rotation ($t(15) = -3.16, p < 0.01$) all showed statistically significant differences. Similar ANOVA results reported in the more detailed analysis given above hold for the fixed-term VI scores. The the main effects of number of terms and iterations are both significant ($F(3,12) = 31.09, p < 0.001$, $F(12,144) = 9.00, p < 0.001$). The interaction between these two factors is significant too ($F(36,144) = 1.97, p < 0.005$).

5.5.4 Other methods for comparing partitions.

The analyses presented in Section 5.4 relied on the Variation of Information as a measure of the difference between partitions. We confirmed that the same results hold using alternative entropy-based measures of the difference between partitions, such as the Adjusted Rand Index and van Dongen's metric (see (Meilă, 2007) for details of these metrics). Given two clusterings C and C' with k and k' clusters each, the contingency table giving the frequency of each pair of cluster labels is $n_{kk'} = |C_k \cap C'_{k'}|$. The Rand Index is based on the idea of counting pairs of points on which the two clusterings agree or disagree. It is defined as

$$R(C,C') = \frac{n_{11} + n_{00}}{n(n-1)/2} \tag{5.5}$$

where n_{11} and n_{00} are the numbers of point pairs that are both in the same cluster and both in different clusters under both C and C', respectively, and n is the total number of data points. The Rand Index was modified to constrain the range between [0,1], yelding the adjusted Rand Index (AR)

$$\begin{aligned} AR(C,C') &= \frac{R(C,C') - E[R]}{1 - E[R]} \\ &= \frac{\sum_{k=1}^{K}\sum_{k'=1}^{K'}\binom{n_{kk'}}{2} - \left[\sum_{k=1}^{K}\binom{n_k}{2}\right]\left[\sum_{k'=1}^{K'}\binom{n'_{k'}}{2}\right]/\binom{n}{2}}{\left[\sum_{k=1}^{K}\binom{n_k}{2} + \sum_{k'=1}^{K'}\binom{n'_{k'}}{2}\right]/2 - \left[\sum_{k=1}^{K}\binom{n_k}{2}\right]\left[\sum_{k'=1}^{K'}\binom{n'_{k'}}{2}\right]/\binom{n}{2}} \end{aligned} \tag{5.6}$$

where n_k and $n'_{k'}$ are the number of data points in clusterings C and C', respectively. For the R and AR metrics, a larger value indicates a closer correspondence between clusterings. The van Dongen (D) metric is defined as

$$D(C,C') = 2n - \sum_{k}\max_{k'} n_{kk'} - \sum_{k'}\max_{k} n_{kk'} \tag{5.7}$$

where a lower value indicates a reduced distance between two clusterings.

Figure 5.5 shows the results using all three metrics, mirroring the analyses in the main paper, that were obtained using VI. As with the original analyses, we computed the distances between random iterated-learning chains and WCS data under the AR and D metrics and compared those values to the distances between the systems produced by individual speakers from the WCS, as well as those produced in our English experiment. All three metrics showed converging results in the rotation analyses (column two) and the comparisons to English (column three). As with the VI analysis, the rotation analyses for both AR and D metrics showed that as the final systems were rotated farther away from the original system, their distance from the WCS increased. Paired t-tests on the AR and D values for no-rotation vs. maximum-rotation (AR: $t(19) = 4.18, p < 0.001$;

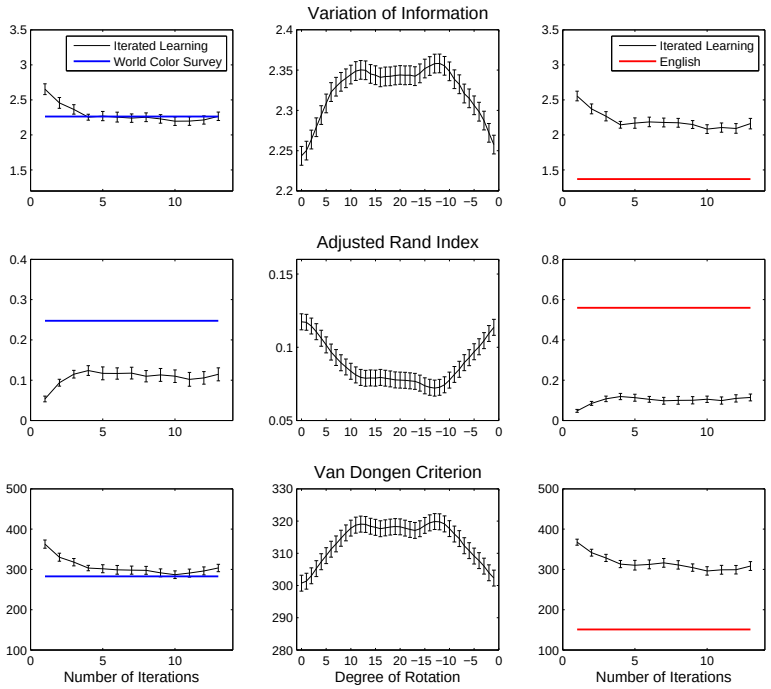

Figure 5.5: Comparing systems of color terms produced by iterated-learning with WCS and English color systems. The distance between systems is assessed using the Variation of Information (VI) in the first row, the Adjusted Rand Index (AR) in the second row, and the Van Dongen metric (D) in the third row. The first column shows the mean distances between systems produced by English-speaking participants in our iterated learning experiment and those seen in the WCS. The second column shows the final systems produced by our laboratory participants rotated along the hue dimension of the WCS array, resulting in a worse fit to the WCS using all three metrics. The third column shows the mean distances between systems produced in the iterated learning experiment and those produced in the English experiment.

D: $t(19) = -5.09, p < 0.001$), no-rotation vs. quarter-rotation (AR: $t(19) = 2.41, p < 0.05$; D: $t(10) = -2.74, p < 0.05$), and no-rotation vs. three- quarter-rotation (AR: $t(19) = 4.15, p < 0.001$; D: $t(19) = -3.92, p < 0.001$) all showed statistically significant differences. Likewise, the distance between systems in iterated learning chains with random initial partitions and systems produced in the English experiment did not get close to the distance between English systems under any of the metrics. A two-samples t- test showed statistically significant differences between the AR ($t(27) = -17.02, p < 0.001$) and D ($t(27) = 9.37, p < 0.001$) scores produced by the final participants in each chain and those for the English systems.

However, the comparison between iterated-learning and WCS results (column one of Figure 5.5) were mixed. All three metrics showed a clear decrease in distance from the WCS data over iterations. A paired t-test on the D values for the initial and final systems in the random chains showed a statistically significant difference ($t(19) = 5.44, p < 0.001$), as did the same test on the AR scores ($t(19) = -3.34, p < 0.005$). The D metric also showed a reduction in distance to a level similar to that between WCS languages. While there was a statistically significant difference in the mean D scores for the iterated learning participants and the WCS data in the final iteration ($t(128) = 3.25, p = 0.0015$), inspection of the figure reveals that this difference is small, and there was no significant differences in iterations 9-11. The AR scores, however, showed a different pattern. The AR between iterated learning participants and WCS speakers never reaches the level of the AR between WCS speakers. A two-samples t-test over the AR scores between the final systems from random chains and WCS and those within WCS speakers yielded a statistically significant difference ($t(128) = -10.45, p < 0.001$).

To understand the different results produced by the AR metric, we examined how the n_{11} and n_{00} counts that contribute to the Rand Index behaved in our data. The n_{11} score indicates the number of pairs of points that are assigned to the same cluster as one another in both clusterings – a measure of the extent to which two chips that share a label in one clustering also share a label in the other clustering. The n_{00} score indicates the number of pairs of points that are assigned to different clusters from one another in both clusterings – a measure of the extent to which two chips that are assigned to different clusters in one clustering are also assigned to different clusters in the other clustering. We divided both n_{11} and n_{00} by the total number of pairs, and then performed the same analyses using these two metrics as we did for the other metrics. The results are shown in Figure 5.6.

The findings for the rotation analysis and comparison to English are similar to those for the other metrics, so we will focus on how the distance to the WCS changed as a function of iteration (column one in Figure 5.6. There are three important things to note. First, there was an increase in the extent to which chips that were given the same term in the WCS were given the same term by our participants over iterations. Second, the score based on n_{11} exceeded the same quantity for the WCS from the second iteration, while the score based on n_{00} never approached that of the WCS. Finally, n_{11} accounts for a far smaller proportion of pairs than n_{00}.

These three observations help to explain why metrics based on the Rand Index are producing poor results. The high n_{11} score and low n_{00} score reflects the fact that the WCS speakers tend to use more terms than the nominal number of terms in the language, while our experimental

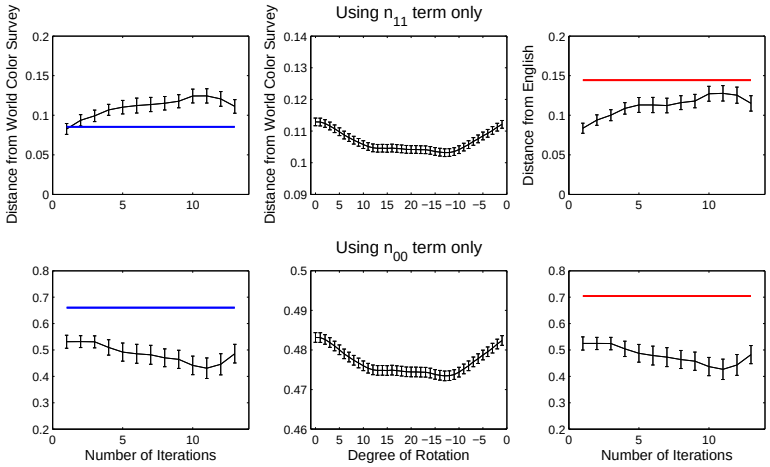

Figure 5.6: Examining the behavior of the proportion of matching pairs n_{11} and mismatching pairs n_{00}. The first column shows the mean distances between systems produced by English-speaking participants in our iterated learning experiment and those seen in the WCS. The second column shows the final systems produced by our laboratory participants rotated along the hue dimension of the WCS array. The third column shows the mean distances between systems produced in the iterated learning experiment and those produced in the English experiment.

participants could use only as many terms as were available in their condition (and sometimes they used fewer). Clustering chips into fewer groups increases the chance of chips being assigned to the same cluster, increasing matching pairs while decreasing mismatching pairs. The absolute values of n_{11} and n_{00} are thus hard to assess against the WCS data. However, the clearest trend here is an increase in the number of matches with the WCS data as the number of iterations increases. This trend is potentially swamped by the larger contribution of n_{00} than n_{11} to the Rand Index simply as a consequence of the larger number of pairs that are assigned to different clusters when the number of clusters is relatively large. Poor handling of cases where the number of clusters differs between clusterings and not providing equal weighting to matching and mismatching pairs are two of the factors that led researchers to criticize measures based on the Rand Index, and motivated the development of the VI metric that we used in our primary analyses (Meilă, 2007).

5.6 Conclusion

Our experiment shows that systems of color terms transmitted in the laboratory by English-speaking participants converge towards forms consistent with the WCS. These results suggest that the learning and perceptual biases of human learners, brought out through cultural transmission, are sufficient to account for the regularities observed in systems of color terms across human languages.

This account provides an alternative to arguments that linguistic universals result from the structure of the coordination problem faced by agents trying to communicate with one another (Steels & Belpaeme, 2005; Baronchelli et al., 2010), and is complementary to analyses based on the idea that universals result from optimally partitioning the space of colors (Shepard, 1992; Yendrikhovskij, 2001; Jameson & D'Andrade, 1997; Regier et al., 2007). Specifically, our results illustrate how systems with such optimal properties can be produced simply through the process of cultural transmission, and locate the forces that lead to such optimal partitions in the human learning and perceptual systems. We anticipate that similar laboratory simulations of cultural evolution will be effective in further elucidating how languages and concepts change through cultural transmission.

Chapter 6

Conclusion: From the self to the world

Control and adaptation are critical for our existence. In one line of research, we show that flexibility and biases go hand-in-hand. People can learn to selectively inhibit their actions, although this ability is subject to biological and psychological constraints. Being flexible is about knowing and using biases, and learning how to limit or modify them. Good control means following a rhythm. In another line of research, we lean heavily on the probabilistic models to discover these biases. As information flows from one person to another, one generation to the next, and between cultures, we can identify biases in our minds.

The probabilistic and functional approaches, combined with new technologies, can be used as converging tools. New technologies open the door for collecting new empirical results, while statistical tools give us a powerful means to get a handle on these data. How do we go about it then? Let's go back to Marr's (1982) three levels: climbing up and down the ladder, we collect the data from the physical world, come up with ideas about mental representations, infer the underlying patterns, and figure out their purposes, and maybe, eventually, contribute to tuning up our system for a larger benefit.

Stepping back a little, we can now draw the link that the process of scientific discovery of the mind is the same thing as an individual's endeavor of figuring out his/her existence in the world. The scientific approach of figuring this out is to collect pieces of information and infer the underlying patterns in the larger scale. This link between the mind and science allows us to join the constant human empirical effort of knowing the world.

Finally, thanks to Judea Pearl for encouraging us, in his commencement speech at University of Toronto (2007), as scientists, to "have respect for the truth, and the audacity to believe that you can find it." Let's hope that our Markov Chain of scientific discovery - constantly collecting pieces of data from the reality, checking against the model of the world we learn from our previous generation, and passing some of our own interpretations along to the next generation - will converge to the truth.

References

Ach N. (1935). Analyse des Willens. In: Abderhalden E, editor. Handbuch der biologischen Arbeitsmethoden (Vol. VI) Berlin: Urban & Schwarzenberg.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on automatic control, 19(6),* 716-723.

Alexander, G., Crutcher, M., & DeLong, M. (1990). Basal ganglia-thalomocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Progress in Brain Research,* 85,119 –146.

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.

Aron AR, Behrens TE, Smith S, Frank MJ, Poldrack RA. (2007). Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *Journal of Neuroscience,* 27(14):3743-52.

Aron, A., & Poldrack, R. A. (2006). Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus. *Journal of Neuroscience, 26,* 2424 2433.

Aron, A. R. & Verbruggen, F. (2008). Stop the presses: dissociating a selective from a global mechanism for stopping. *Psychological Science,* 19, 1146-53.

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, 14, 208-215.

Band, G. P. H. , van der Molen, M. W., & Logan, G. D. (2003). Horse-race model simulations of the stop-signal procedure. *Acta Psychologia*, 112, 105-142.

Bangerter, A. (2000). Transformation between scientific and social representations of conception: The method of serial reproduction. *British Journal of Social Psychology*, 39, 521–535.

Bappu, A., & Griffiths, T. L. (2009). Iterated learning and the cultural ratchet. In *Proceedings of the 31th annual conference of the cognitive science society.*

Bargh, J. A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin*, 126, 925–945.

Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences USA, 107*, 2403-2407.

Barrett, J., & Nyhof, M. (2001). Spreading nonnatural concepts: The role of intuitive concep- tual structures in memory and transmission of cultural materials. *Journal of Cognition and Culture, 1*, 69-100.

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

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley, CA: University of California Press.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-foraction in anterior cingulate cortex. *Nature,* 402, 179–181.

Botvinick, M. M., Carter, C. S., Braver, T.S., Barch, D.M., & Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624–652.

Boucher, L., Stuphorn, V., Logan, G. D., Schall, J. D., & Thomas J. Palmeri Stopping eye and hand movements: Are the processes independent? *Attention, Perception, & Psychophysics*, 69(5), 785-801.

Box, G. E. P., & Jenkins, G. M. (1994). *Time series analysis: Forecasting and control*. Upper Saddle River, NJ: Prentice Hall.

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: cross-cultural evi- dence for recall of counter-intuitive representations. *Cognitive Science*, *25*, 535-564.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433-436.

Broadbent, D. E. (1958). *Causality: Models, reasoning, and inference*. New York: Pergamon Press.

Bullmore, E., Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Review Neuroscience,* 10, 186-198.

Coxon, J. P., Stinear, C. M., &Byblow, W. D. (2007). Selective inhibition of movement. *Journal of Neurophysiology*, 97, 2480-9.

Cournot, A. A. (1843). *Exposition de la theorie de chances et des probabilities.* Paris: Hachette.
De Jong, R., Coles, M. G., &Logan, G. D. (1995). Strategies and mechanisms in nonselective and selective inhibitory motor control. *Journal of Experimental Psycholology: Human Perception and Performance*, 21, 498-511.

Davidoff, J., Davies, I., & Roberson, D. (1999). Color categories of a stone-age tribe. *Nature, 398*, 203-204.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B, 39.*

Derrfuss, J., Brass, M., & von Cramon, D. Y. (2004). Cognitive control in the posterior frontolateral cortex: Evidence from common activations in task coordination, interference control, and working memory. *Neuroimage,* 23, 604–612.

Dowman, M. (2007). Explaining color term typology with an evolutionary model. *Cognitive Science*, *31*(1), 99-132.

Dowman, M. (2009). Evolution of basic color terms. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language evolution and the brain* (p. 109-139). Hong Kong: City University of Hong Kong Press.

Duque, J., Lew, D., Mazzocchio, R., Olivier, E., & Ivry, R. B. (2010). Evidence for two concurrent inhibitory mechanisms during response preparation. *Journal of Neuroscience*, 30, 3793-802.

Duque, J., & Ivry, R. B. (2009) Role of corticospinal suppression during motor preparation. *Cerebral Cortex*, 19, 2013–2024.

Duque, J., Mazzocchio, R., Dambrosia, J., Murase, N., Olivier, E., & Cohen, L. G. (2005). Kinematically specific interhemispheric inhibition operating in the process of generation of a voluntary movement. Cereb Cortex 15:588 –593.

Duque, J., Murase, N, Celnik, P., Hummel, F., Harris-Love, M., Mazzocchio, R., Olivier, E., & Cohen, L. G. (2007). Intermanual differences in movementrelated interhemispheric inhibition. *Journal of Cognitive Neuroscience*, 19, 204 –213.

Eriksen, B. A., & Eriksen, C. W. Effects of noise letters upon identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16, 143–149

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630-633.

Fox, M., & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Review Neuroscience, 8*, 700-11.

Fox, M. D., Snyder, A. Z., Zacks, J. M., & Raichle, M. E. (2006). Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature Neuroscience*, 9, 23–25.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Golland, Y., Bentin, S., Gelbard, H., Benjamini, Y., Heller, R., Nir, Y., Hasson, U., & Malach, R. (2006). Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. Cerebral Cortex, 17, 766-777.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for revealing inductive biases. *Cognitive Science, 32*, 68-107.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty- Seventh Annual Conference of the Cognitive Science Society* (p. 827-832). Mahwah, NJ: Erlbaum.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science, 31*, 441-480.

Hesselmann, G., Kell, C. A., Eger, E., Kleinschmidt, A. (2008) Spontaneous local variations in ongoing neural activity bias perceptual decisions. *Proceedings of the National Academy of Sciences USA*, 105, 10984–10989.

Heider, E. (1972). Probabilities, sampling, and ethnographic method: The case of dani color names. *Man, 7*, 448-466.

Hemmer, P., & Steyvers, M. (2008). A bayesian account of reconstructive memory. In

Proceedings of the 30th annual conference of the cognitive science society.

Hume, D. (1748/1910). *An enquiry concerning human understanding*. Clacton on SeaEssex: Collier and Son Ltd.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judg- ment? *Journal of Experimental Psychology: General,* 220-241.

James, W. (1890). *Principles of psychology*. New York: Holt.

Jameson, K., & D'Andrade, R. G. (1997). It's not really Red, Green, Yellow, Blue: An inquiry into cognitive color space. In C. L. Hardin & L. Maffi (Eds.), *Color categorization in thought and language.* Cambridge: Cambridge University Press.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*.

Kantowitz, B. H. (1974). Double stimulation. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 83-131). Hillsdale, NJ: Erlbaum.

Kay, P., Berlin, B., Maffi, L., & Merrifield, W. R. (1997). Color naming across languages. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language.* Cambridge, UK: Cambridge University Press.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. Stanford: Center for the Study of Language and Information.

Kay, P., Berlin, B., & Merrifield, W. R. (1991). Biocultural implications of systems of color naming. *Journal of Linguistic Anthropology*, *1*, 12-25.

Kay, P., & Maffi, L. (1999). Color appearance and the emergence and evolution of basic color lexicon. *American Anthropologist*, *101*, 743-760.

Kay, P., & McDaniel, C. (1978). The linguistic significance of the meanings of basic color terms. *Language*, *54*, 610-646.

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, *100*, 9085-9089.

Kahneman, D. (197 3). *A ttention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Keele, S. W., & Neill, W. T. (1978). Mechanisms of attention. In E. C. Carterette & P. Friedman (Eds.), *Handbook of perception* (Vol. 9, pp.3-47). New York: Academic Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102-110.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*, 5241-5245.

Kramer, A. F., Humphrey, D. G., Larish, J. F., Logan, G. D., & Strayer, D. L. (1994). Aging and inhibition: Beyond a unitary view of inhibitory processing in attention. *Psychology and Aging*, 9, 491–512.

Laufs, H., Krakow, K., Sterzer, P., Eger, E., Beyerle, A., Salek-Haddadi, A., &. Kleinschmidt, A. (2003). Electroencephalographic signatures of attentional and cognitive default

modes in spontaneous brain activity fluctuations at rest, *Proceedings of the National Academy of Sciences USA*, 100, 11053–11058.

Lappin, J. S., & Eriksen, C. W. (1966). Use of a delayed signal to stop a visual reaction-time response. *Journal of Experimental Psychology, 72*, 805-811.

Levitt, H. (1971). Transformed up-down methods in psycboacoustics. *The Journal of the Acoustical Society of America*, 49, 467-477.

Liberman, A. M., Cooper, F. S., shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431-461.

Lindquist, M. A., & Gelman, A. (2009). Correlations and multiple comparisons in functional imaging - a statistical perspective. *Perspectives in Psychological Science, 4*, 310-313.

Logan, G. D. (1994). On the ability to inhibit thought and action: a users' guide to the stop signal paradigm. In D. Dagenbach & T. H. Car (Eds.), *Inhibitory processes in attention, memory, and language* (p. 189-239). San Diego: Academic Press.

Logan, G. D. & Burkell, J. (1986). Dependence and independence in responding to double stimulation: A comparison of stop, change, and dual-task paradigms. *Journal of Experimental Psychology: Human Perception and Performance,* 12(4), 549-563

Logan, G. D., & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review, 91*, 295-327.

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit simple and choice re- action time responses: a model and a method. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 276-91.

Majid, D. S. A., Cai, W., George, J., Verbruggen, F., & Aron, A. R. (2011). Transcranial magnetic stimulation reveals dissociable mechanisms for global versus selective corticomotor suppression underlying the stopping of action. *Cerebral Cortex*.

Malsburge, C. von der. (1994). The correlation theory of brain function. In E. Domany, J. L. V. Hemmen, & K. Schulten (Eds.), *Models of neural networks ii* (p. 95-119). Berlin: Springer.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound mod- els in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 128-148.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analalysis, 98,* 873–895.

Mesoudi, A. (2007). Using the methods of experimental social psychology to study cultural evolution. *Journal of Social, Evolutionary, and Cultural Psychology, 1*, 35-38.

Mesoudi, A., Whiten, A., & Laland, K. N. (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences, 29*, 329-383.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and Strcuture of Behavior*. New York: Holt, Rinehart & Winston.

Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50, 381– 425.

Norris, J. R. (1997). *Markov chains*. Cambridge, UK: Cambridge University Press.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9, 97–113.

Osman, A., Kornblum, S., & Meyer, D. E. (1986). The point of no return in choice reaction time: Controlled and ballistic stages of response preparation. *Journal of Experimental Psychology: Human Perception and Performance*, 12, 243-258.

Osman, A., Kornblum, S., & Meyer, D. E. (1990). Does motor programming necessitate response execution? *Journal of Experimental Psychology: Human Perception and Performance*, 16, 183-198.

Pearl, J. (2000). *Perception and communication*. Cambridge, UK: Cambridge University Press. Poldrack, R.A. Sabb, F. W., Foerde, K., Tom, S. M., Asarnow, R. F., Bookheimer, S. Y. & Knowlton, B. J. (2005). The neural correlates of motor skill automaticity. *Journal of Neuroscience*, 25, 5356–5364

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval Processes in Recognition Memory. *Psychological Review*, 83(3), 173-89.

Rieger, M., & Gauggel, S. (1999). Inhibitory after-effects in the stop signal paradigm. *British Journal of Psychology, 90*, 509–518.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences, 104*, 1436-1441.

Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.

Ridderinkhof, K. R. (2002). Activation and suppression in conflict tasks: Empirical clarification through distributional analyses. In B. H. W. Prinz (Ed.), *Mechanisms in perception and action* (p. 494-519). Oxford: Oxford University Press.

Ridderinkhof, K. R., Band, G. P. H. & Logan, G. D. (1999) A study of adaptive behavior: Effects of age and irrelevant information on the ability to inhibit one's actions. *Acta Psychologica.* 101, 315–337.

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

Sailor, K. M., & Antoine, M. (2005). Is memory for stimulus magnitude Bayesian? *Memory and Cognition, 33*, 840-851.

Salinas, E., & Sejnowski, T. J. (2001). Correlated neuronal activity and the flow of neural infor- mation. *Nature Review Neuroscience, 2*, 539-550.

Schneider, W., Shiffrin, R. M. (1977). Controlled and automatic human information processing. I. Detection, search and attention. *Psychological Review*, 84, 1–66.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of statistics, 6(2)*, 461-464.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A.L. & Greicius, M. D. Dissociable intrinsic connectivity networks for salience processing and executive control, *Journal of Neuroscience*, 27, 2349–2356.

Sejnowski, T. J. (1981). Skeleton filters in the brain. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (p. 189-212). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Serrien, D. J. & Swinnen, S. P. (1997). Coordination constraints induced by effector combination under isofrequency and multifrequency conditions, *Journal of Experimental Psychology: Human Perception and Performance*, 23, 1493–1510.

Shepard, R. N. (1992). The perceptual organization of colors. In J. B. et al. (Ed.), *The adapted mind* (p. 495-532). Oxford: Oxford University Press.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifi- cations. *Psychological Monographs, 75*. (13, Whole No. 517)

Sohn, Y. H., Jung, H. Y., Kaelin-Lang, A., Hallett, M. (2003). Excitability of the ipsilateral motor cortex during phasic voluntary hand movement. *Experimental Brain Research*, 148, 176-85.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptual categories through language: A case study for colour. *Behavioral and Brain Sciences, 28*, 469-489.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*, 881-911.

Swinnen, S. P., Jardin, K., Neulenbroek, R., Dounskaia, N., & Hofkens-Van Den Brandt M. (1997). Egocentric and allocentric constraints in the expression of patterns of interlimb coordination, *Journal of Cognitive Neuroscience*, 9, 348–377.

Verbruggen, F., & Logan, G.D. (2008a). Aftereffects of goal shifting and response inhibition: A comparison of the stop-change and dual-task paradigms. *Quarterly Journal of Experimental Psychology*, 61, 1151–1159.

Verbruggen, F., & Logan, G. D. (2008b). Automatic and controlled response inhibition: associative learning in the go/no-go and stop-signal paradigms. Journal of Experimental Psychology: General,137(4), 649-72.

Verbruggen, F., Liefooghe, B., Notebaert, W., & Vandierendonck, A. (2005). Effects of stimulus-stimulus compatibility and stimulus-response compatibility on response inhibition. *Acta Psycholgia*, 120(3), 307-26.

Verbruggen, F., Liefooghe, B. & Vandierendonck, A. (2004). The interaction between stop signal inhibition and distractor interference in the flanker and the Stroop task. *Acta Psychologica*, 116, 21–37.

Verbruggen, F., & Logan, G. D. (2008c). Long-term aftereffects of response inhibition: Memory retrieval, task goals, and cognitive control. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 1229-1235.

Verbruggen, F., & Logan, G. D. (2008d). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Science*, 12, 418-424.

Verbruggen, F., Logan, G. D., Liefooghe, B., & Vandierendonck, A. (2008). Short-term aftereffects of response inhibition: Repetition priming or between-trial control adjustments. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 413–426.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274-290.

Wiener, N. (1949). *The extrapolation, interpolation and smoothing of stationary time series*. New York: John Wiley and Sons, Inc.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman, R. Davies, & J. Beatty (Eds.), *Varieties of attention* (pp. 63-102). New York: Academic Press

Xue, G., Aron, A.R., Poldrack, R.A., (2008). Common neural substrates for inhibition of spoken and manual responses. *Cerebral Cortex*, 18, 1923-32.

Yamaguchi, M., Logan, G. D., & Bissett, P. G. (2011). Stopping while going! Response inhibition does not suffer dual-task interference. Journal of Experimental Psychology: Human Perception and Performance.

Yendrikhovskij, S. N. (2001). Computing color categories from statistics of natural images. *Journal of Imaging Science and Technology, 45*, 409-417.

Yin, H. H., & Knowlton, B.J. (2006). The role of the basal ganglia in habit formation. *Nature Review Neuroscience*, 7, 464–476.