THE UNIVERSITY
of EDINBURGH

Inductive evolution: cognition, culture, and regularity in language

Vanessa Ferdinand

Doctor of Philosophy
University of Edinburgh
2015

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Vanessa Ferdinand*)

Abstract

Cultural artifacts, such as language, survive and replicate by passing from mind to mind. Cultural evolution always proceeds by an inductive process, where behaviors are never directly copied, but reverse engineered by the cognitive mechanisms involved in learning and production. I will refer to this type of evolutionary change as inductive evolution and explain how this represents a broader class of evolutionary processes that can include both neutral and selective evolution.

This thesis takes a mechanistic approach to understanding the forces of evolution underlying change in culture over time, where the mechanisms of change are sought within human cognition. I define culture as anything that replicates by passing through a cognitive system and take language as a premier example of culture, because of the wealth of knowledge about linguistic behaviors (external language) and its cognitive processing mechanisms (internal language). Mainstream cultural evolution theories related to social learning and social transmission of information define culture ideationally, as the subset of socially-acquired information in cognition that affects behaviors. Their goal is to explain behaviors with culture and avoid circularity by defining behaviors as markedly not part of culture. I take a reductionistic approach and argue that all there is to culture is brain states and behaviors, and further, that a complete explanation of the forces of cultural change can not be explained by a subset of cognition related to social learning, but necessarily involves domain-general mechanisms, because cognition is an integrated system. Such an approach should decompose culture into its constituent parts and explore 1) how brains states effect behavior, 2) how behavior effects brain states, and 3) how brain states and behaviors change over time when they are linked up in a process of cultural transmission, where one person's behavior is the input to another.

I conduct several psychological experiments on frequency learning with adult learners and describe the behavioral biases that alter the frequencies of linguistic variants over time. I also fit probabilistic models of cognition to participant data to understand the inductive biases at play during linguistic frequency learning.

Using these inductive and behavioral biases, I infer a Markov model over my empirical data to extrapolate participants' behavior forward in cultural evolutionary time and determine equivalences (and divergences) between inductive evolution and standard models from population genetics. As a key divergence point, I introduce the concept of non-binomial cultural drift, argue that this is a rampant form of neutral evolution in culture, and empirically demonstrate that probability matching is one such inductive mechanism that results in non-binomial cultural drift. I argue further that all inductive problems involving representativeness are potential drivers of neutral evolution unique to cultural systems. I also explore deviations from probability matching and describe non-neutral evolution due to inductive regularization biases in a linguistic and non-linguistic domain. Here, I offer a new take on an old debate about the domain-specificity vs -generality of the cognitive mechanisms involved in language processing, and show that the evolution of regularity in language cannot be predicted in isolation from the general cognitive mechanisms involved in frequency learning. Using my empirical data on regularization vs probability matching, I demonstrate how the use of appropriate non-binomial null hypotheses offers us greater precision in determining the strength of selective forces in cultural evolution.

Acknowledgements

First and foremost, I would like to acknowledge my supervisor, Simon Kirby, for the immeasurable source of inspiration and support (in all aspects of life: academic, existential, and musical) that he has been to me over the past four years. A while back we joked about over-the-top acknowledgements and I floated the idea of including 200 blank pages right here in representation of what this thesis would have been without him. That's not say he wrote this thing, but he certainly provided the keystone against which I could lay my ideas. Working with him is a wonderfully synergistic experience.

Second and secondmost, I must acknowledge my other supervisor, Kenny Smith, who taught me that British humor happens when someone is mean to someone else, and then they both laugh. There are many problems in life that can be solved by just having someone tell you the answer, and I am greatful to have had Kenny around to provide so many of these answers. Sitting in his office with him at the whiteboard, everything made sense, and I supposed that's why I returned with the same questions so many times.

There are a great many others who have contributed profoundly to the ideas in this thesis. I would like to thank, wholeheartedly:

Simon DeDeo and Tom Griffiths, two of my most admired academics, whom I have been privileged to unofficially call my supervisor for some brief and fruitful months in Sante Fe and Berkeley.

Luke Maurits, for the slew of enlightening conversations, his excellent comments on Chapter 5, and for co-founding with me "Linear Algebra Club, Berkeley, Spring of 2013", membership: 2. Andy Wedel, for being an all-around inspiring human being, for cooking me pasta in Tesuque, and for being such a reliable dance partner at conferences. Olga Feher, for keeping life at the LEC fun with all of the

hilarious circumstances better left unprinted. Evandro Ferrada, for geeking out with me about drift on complex landscapes and for his friendship, kindness, and support on every one of my trips to Santa Fe. And Michael Dunn, Bart De Boer, Monica Tamariz, Thom Scott-Phillips, Jim Hurford, Carol Paddon, Amy Perfors, Nikolaus Ritt, Tessa Verhoef, Andrea Ravignani, Florencia Reali, Jennifer Culbertson, Marieke Schouwstra, Jelle Zuidema, Vikram Vijayaraghavan, and Ryan James, for the many exciting exchanges about language evolution, regularization, and information theory.

Caroline Kamps, for the excellent work she did on the design and implementation of Experiment 5 in this thesis. Co-supervising her was the highlight of my time at the LEC.

The handful of other PhD students in this world I have been fortunate enough to have had close collaborations with are: Sabine van der Ham, Katrien Beuls, Enrico Sandro Colizzi, Jasmeen Kanwal (here's to all of the adventures we've had in foreign lands on the coat-tails of conferences, summer schools, and such), and Bill Thompson (my intellectual soul mate who knows me well enough to believe that the sharpie markered "tattoo" of Bayes Rule on my bicep when I returned from Berkeley was real).

All of the PhD students at the Language Evolution and Computation Research Unit: Sean Roberts, Hannah Cornish, Rachael Bailes, Christine Cuskley, Justin Quillinan, Justin Sulik, Keelin Murray, Marton Soskuthy, Bill Thompson, Matt Spike, Carmen Saldaña, Mark Atkinson, Yasamin Motamedi-Mousavi, Catriona Silvey, Alan Nielsen, James Thomas, James Winters, Jon Carr, and last but not least, Kevin Stadler (who is on call to bind and submit these pages as are, in the event that I pass out unrecoverably from mental fatigue). And special thanks again to Carmen Saldaña, Yasamin Motamedi-Mousavi, and Catriona Silvey for convincing me to socially copy that high-frequency lunch-eating trait people exhibit and for being such a reliable source of support as I was writing up.

Luc Steels and all of the participants at the International Summerschool on Agent-based Models of Creativity in Cortona, 2013. And all of the participants at the 2012 Complex Systems Summer School in Santa Fe and that inspiring hilltop institute to which I will be returning, as soon as this thesis is deemed acceptable.

The only reason I do the work that I do is so I can get to talk about it with everyone listed above, I'm for serious.

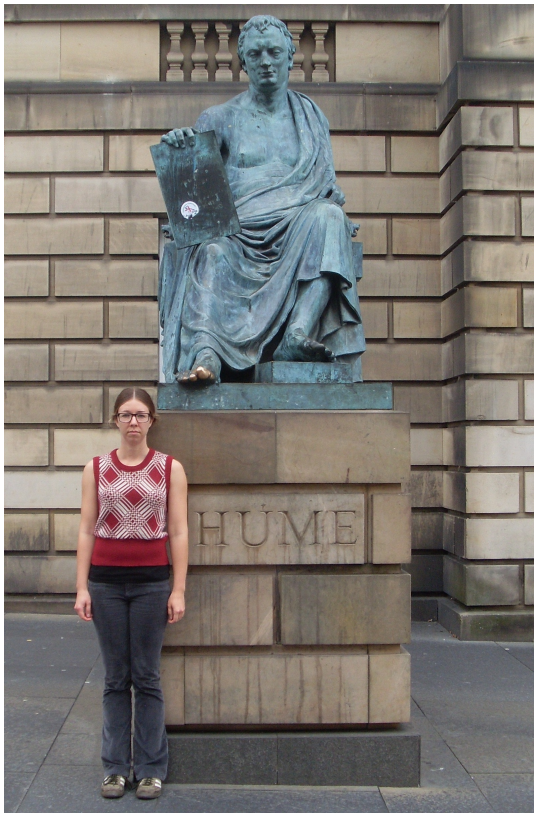
I would also like to thank the funding bodies who so generously supported my research: the University of Edinburgh's College Studentship, the SORSAS award, and the Engineering and Physical Sciences Research Council.

Perhaps most importantly, were all of the people who supported me on the home front. Here, I would like to thank my home skillets Mia Browne and Matthew Flinn (featured on p. 133) for keeping me watered and well-fed. Emma Dunmore, for her constant support, garnished with banjo duets and walks on Blackford Hill. My pals Caro Kemp, Rhiannon Sims, Rea Cris, and Kathleen Messer, for all the times we've shared together (the good, the bad, the rank, the radge) and who I thank, on behalf of my thesis, for kicking me out of the band. Gregory Myles, my roommate and parakeet-baby-daddy, who stepped aside from his fine Dundee diction to kindly proof-read so much of my final draft. Lindsay Hunter, for grounding me in the relentless technical drumcore of the Beltane Fire Society, which turned out to be a seasonal cult, but whatever. Alessia De Stefani, who discovered me in this world as a someone about to start writing up her PhD thesis and stuck by me to the end. However she managed that, I may never understand. And finally I would like to thank my family, especially Jane Ferdinand, Laura Ferdinand, Alejandro Fernandez, Beverly Ball, and Marjorie Van Halteren, for supporting the academic adventures which took me one step farther away from home every time.

To the dear and mysterious reader,
who hasn't supervised or examined this thesis,
these pages are for you

and to David Hume,

who narrowly saved me from postgraduate study in epistemology
with the lure of cognitive science
and then perhaps Edinburgh.



"what the vulgar call chance is nothing but a secret and conceal'd cause."
A Treatise of Human Nature, 1739

Contents

Chapter 1

Cultural evolution is inductive evolution

1.1 Introduction

Cultural artifacts, such as language, music, and technology, survive and replicate by passing from one mind to another. In the absence of cognition, which can perceive, learn, and produce behaviors for others to see, culture would not exist.

This thesis is concerned with the type of cultural variation found in the world and how it got there. The term *cultural variation* provokes images of the diverse cuisines and clothing found around the world, the different types of homes people build for themselves and live in, the various political systems and forms of governance that structure societies, and the thousands of languages that humankind uses to communicate with one another. These things vary between societies, within societies, and even within individuals and constitute synchronic variation: the variation we see in one snapshot, at a particular moment in time. We also know very well that culture changes over time. You don't wear the exact same type of clothing that your great grandparents wore, the slang you use today is (hopefully) different than the slang you used a decade ago, and your computer probably doesn't weigh 20 kilos anymore. These changes over time constitute diachronic variation and are closely intertwined with synchronic variation, which fuels the possibilities of change at any given moment.

To understand how and why culture changes, we need to understand what happens at the locus of cultural change. But what is the locus of cultural change? Where do manuscripts accumulate typos, where do stone tools develop smaller, finer blades, and where do objects get new names? The answer to this question is cognition. Cultural variants are never directly copied, but reproduced through

a cycle of perception, processing, and production. For a stone tool to be copied, it needs someone to observe its properties, infer how those properties came to be, and then implement them in another piece of stone. This is a process of reverse engineering (Kirby, 2013). Cultural variants can either be reverse engineered without error, leading to perfect transmission fidelity, or with error, leading to change. And these errors can either introduce new variants, or turn a variant into another existing variant.

Some things are easier to learn or produce than others, and those that are difficult to replicate with high fidelity will fall out of the pool of cultural variation over time. Variants that are easier to reproduce with high fidelity, or are the product of a type of error that is consistently made, will increase in number over time. It seems likely that most of the errors humans make are not truly random. Cognitive systems contain biases and the errors that we make while trying to copy or learn things are often structured in interesting ways due to a variety of interesting reasons. Distributions of variants are shaped by these biases over time and therefore, can be explained in terms of these biases. When we understand how these shaping forces operate, then we understand how the variants came to be distributed as they are, and why one particular distribution exists, rather than some other distribution.

This thesis aims to explore how cognitive biases shape cultural evolution at a very basic level, by looking at the biases we employ when learning about the frequencies of events and objects in our environment, because distortions to these frequencies directly affect the distribution of cultural variants over time. I will focus on a particular type of culturally transmitted entity, language, and detail some of the main cognitive biases that affect the learning and production of word frequencies and how these biases combine to drive the gradual change of linguistic systems over time. Along the way, I will ask, what types of evolutionary change do cognitive biases lead to? Can human frequency learning mechanisms support cultural drift as well as selection? What is special about cultural evolutionary processes that might not have parallels in genetic evolution? And how much information do cognitive biases, and the behavior of individual learners, carry about the long-term evolutionary trajectories that culturally transmitted behaviors take?

1.2 Defining culture

If we want to be explicit about the forces of change that shape culture, we need to be explicit about what we mean by culture. Therefore, the first task at hand is to lay out a workable definition of culture conducive to an evolutionary research agenda.

The vast majority of existing definitions of culture are *ideational* (Kroeber and Kluckhohn, 1952; Keesing, 1974), meaning that they define culture exclusively in terms of ideas. In a comprehensive review, Keesing (1974) grouped definitions of culture into four broad classes: one materialistic view and three ideational views. The first (materialistic) class encompasses *adaptationist* theories of culture. These definitions treat culture as the material ways of being and behaving that allow individuals a better fit to their environment, above and beyond the adaptations that genes are capable of providing. In these views, the diversity of cultures are often seen as the result of fine-tuned adaptations to the different ecologies and geographies that human populations inhabit. For example:

> Culture is all those means whose forms are not under direct genetic control ... which serve to adjust individuals and groups within their ecological communities (Binford, 1968, p. 323).

Keesing's three ideational classes discuss culture as *cognitive systems* (in which cultures are systems of knowledge), *structural systems* (in which cultures are world views with arbitrary, as opposed to natural, order), and *symbolic systems* (in which cultures are systems of shared symbols and meanings). In these definitions, culture is all about knowledge, ideas, beliefs, and values that people individually or collectively hold.

Since Keesing's review, there have been sweeping developments in the study of cultural evolution, extending beyond the field of anthropology. Some of these approaches, which Keesing would likely classify under this adaptationist category, fall largely under the field of sociobiology and approach culture in terms of *niche construction* (e.g. Odling-Smee et al., 2003), *gene-culture coevolution* (e.g. Cavalli-Sforza and Feldman, 1973; Hinton and Nowlan, 1987; Feldman and Laland, 1996; Chater et al., 2009), and concepts of an *extended phenotype* (e.g. Dawkins, 1999). These approaches treat culture as a set of phenotypic traits, subject to natural selection on genomes, that increase an individual's adaptive fit with their environment, and may or may not alter the selection pressures on genomes.

Other new lines of research in *social learning* (e.g. Boyd and Richerson, 1985) and *iterated learning* (e.g. Kirby, 2001) are also adaptational in nature, but view

culture as an evolving system in its own right, without recourse to natural selection on genes to explain cultural change. These approaches treat adaptation the other way around: not how culture increases individuals' adaptive fit to their environment, but rather, how culture itself evolves to fit humans well, by adapting to human minds and social systems.

I will review these new evolutionary approaches more thoroughly in the next section, but what I want to point out here is that all of these approaches to culture, despite being adaptationist in some sense, are largely dominated by ideational definitions of culture. For example:

> Culture is information capable of affecting individuals' phenotypes which they acquire from other conspecifics by teaching or imitation. (Boyd and Richerson, 1985, p. 33)

> Culture is information that is acquired from other individuals via social transmission mechanisms such as imitation, teaching, or language. (Mesoudi, 2011, p. 2)

These definitions attribute to culture only socially-acquired information, but not behaviors. Social learning theorists explicitly exclude behaviors and behavioral artifacts from their definition of culture because they want to use culture as a concept for explaining behavior (Mesoudi, 2011). Including behavior in the definition of culture would 1) make this explanation circular and 2) obscure a direct investigation of culture because there are other causes of behavior besides culture (Cronk, 1999). Mesoudi (2011) states that there are two forms of information, in addition to culture, which affect behavior: information that is acquired *genetically* and information that is acquired by *individual learning*.

In cultural evolution research, the adaptational aspect of culture is necessarily embraced. However, the materialistic definition that traditionally accompanies it has been rejected. The preference for ideational definitions among cultural evolution researchers may be due to the rise of the gene-selectionist perspective in evolutionary biology that took hold in the late 1970's, in the work of Bill Hamilton, John Maynard-Smith, Robert Trivers and George Williams, and crystallized in Richard Dawkins' book *The Selfish Gene* (1976). This viewpoint privileges genetic information as the most important aspect of an evolving system, as a sort of instructional code that ultimately determines an organism's phenotype. Mesoudi (2011) makes this parallel clear:

> Whereas genetic information is stored in sequences of DNA base pairs, culturally transmitted information is stored in the brain as patterns of neural connections ... And whereas genetic information is expressed

> as proteins and ultimately physical structures such as limbs and eyes, culturally acquired information is expressed in the form of behavior, speech, artifacts, and institutions. (p. 3)

Culture, as patterns of neural connections, does carry information about behaviors, but behaviors also carry information about culture. Patterns of neural connections can never be directly transmitted between individuals: the cultural information that they encode can only be inferred from the behaviors of other individuals. If behaviors did not carry information about culture, cultural information could not be acquired by others. Likewise, behaviors can never be directly transmitted either: they will only be culturally inherited if they are perceived, processed, and produced by others. Therefore culture, in a broader sense, seems to have two distinct phases in its life cycle; as information that resides in cognition and information that resides in behaviors and concrete artifacts.

This two-phase cycle is recognized by Hurford (2003) in relation to language, which is perhaps the most prominent subset of human culture. "Because language emerges from the interaction of minds and data, linguistics *must* concern itself with both phases in this life-cycle" (Hurford, 2003, p. 51). In light of this view, Hurford defines language as neither its internal form (I-Language) or external form (E-Language)[1] , but as "their dynamic interaction" and states that "Defining a language in this way is hardly elegant, but (a) it recognizes the essential interdependence of the two phases of language, I-Language and E-Language, and (b) it avoids an arbitrary privileging of one phase over another." (Hurford, 2002)

In line with Hurford, and the general sentiment of researchers in language evolution, I think that a definition of culture should embrace "the spiraling interaction of the two phases" (Hurford, 2003, p. 51), privilege neither its cognitive nor behavioral information content, and be sufficiently broad to include the concept of culture as an emergent phenomenon, while also including culture as cognitive and behavioral information specifically. This brings me to my first stab at defining culture, and will be the definition that I will operationalize throughout this thesis:

> **Culture** is anything that replicates by passing through cognition.

This definition probably leaves much to be desired, but in its application, much can be learned. First, according to this definition, both behaviors and brain states constitute culture, and this may be argued to reinstate the circular reasoning that a purely ideational, or purely materialistic, definition of culture

[1] See Chomsky (1965) for the origin of this terminology.

avoids. However, if we don two lenses, one of reductionism and one of a cognitive scientist, we can break the system of cultural evolution into three constituent parts: 1) how brain states affect behaviors, 2) how behaviors affect brain states, and 3) how cultural transmission links brain states and behaviors into an inheritance cycle that makes cultural evolution possible. These three areas of inquiry outline a framework for cultural evolution research and avoid circularity in explanation by focusing on the processes that underpin cultural evolution, rather than culture itself. Secondly, and importantly, this view does not privilege a particular form or source of information that may affect brain states or behaviors. In particular, it does not make a distinction between culturally-acquired information, individually-acquired information, or innate information. All of these sources of information affect behaviors and a complete description of how a culturally-transmitted behavior, like language, changes over time requires an understanding of all of these sources of constraints. This point will be made concrete in the next chapter, when we see how domain-specific and domain-general biases affect frequency learning in language, and again in Chapter 6 when we see how neither one of these biases can, on its own, predict the ultimate form that a distribution of linguistic variants takes after several generations of learners.

1.3 Cultural transmission

Two predominant frameworks for understanding cultural transmission in the cultural evolution literature are *iterated learning* and *social learning*. Kirby et al. (2014) provide a definition for the first framework:

> **Iterated learning** is the process by which a behaviour arises in one individual through induction on the basis of observations of behaviour in another individual *who acquired that behaviour in the same way.*

Iterated learning originally took foot in the form of agent-based simulations of language evolution as a way to understand the dual contributions that both biological and cultural evolution could have in explaining the structure of human language (Hurford, 1989). This sparked interest in the non-biological forces that shape communication systems, such as interaction and negotiation (Steels, 1999, 2003; Vogt, 2005) and learning biases (Oliphant, 1999; Smith, 2002), and constraints in the transmission process itself, such as the learning bottleneck (Kirby, 2001, 2000, 2002; Brighton, 2002; Zuidema, 2003). A potentially infinite amount of unique utterances are possible from the grammar of human languages, however these grammars are acquired from only a finite amount of linguistic input

(Chomsky, 1965). The learning bottleneck defines the size of that finite data set and iterated learning research showed it to be the driving force behind the emergence of compositional languages. Assuming an initial language that is unstructured (i.e. a long list of idiosyncratic utterances that do not reuse parts of other utterances), when the learning bottleneck is wide and agents observe nearly all possible utterances, they learn the language holistically and the language remains unstructured. However, when the learning bottleneck is small and they see a sufficiently small subset of the utterances, generalizations are formed, subcomponents get re-used, and the language gradually becomes structured over several generations of learners. This constitutes an adaptive process of the language itself to the learning algorithms that process and transmit it. In this sense, cognition provides a transformative copying process that can produce directional change (from unstructured to structured) in a language over time.

On the mathematical modeling front, Nowak and colleagues showed that iterated learning could be recast in terms of the replicator dynamics models of biological evolution (Nowak and Komarova, 2001; Nowak et al., 2001; Komarova et al., 2001; Nowak et al., 2002) and explored the joint contribution of biological fitness and cultural transmission on composition of a population in terms of how many end up speaking the same language. Subsequently, (Griffiths and Kalish, 2007) developed the first mathematical characterization of iterated learning as a purely cultural phenomenon, without recourse to natural selection or the biological fitness of culturally-acquired traits. They made the learning biases explicit by using Bayesian agents, which have clear, quantifiable prior biases. The main result here was that the the population's composition of language types, after many generations of learners, will come to mirror the prior probability that each agent assigns to each language type. This means that the cultural transmission mechanism eventually leads languages to mirror the minds of those who learn them and is known as *convergence to the prior*. Follow-up work by Kirby et al. (2007) identified different Bayesian models in which this convergence result did not hold, and cultural transmission itself seemed to add something to the story.

More recently, the iterated learning research paradigm has taken a decidedly experimental turn. These experiments take the form of traditional psychology experiments where participants are trained on input data of a particular type and then tested on it, however this testing data is then used as the training data for the next participant, and so on. This simulates cultural transmission in the same way as the earlier agent-based models do, but implemented in a population of human learners (with the added benefit of not needing to program a learning algorithm). These experiments have been used in two broad ways. First, they have been used

to confirm the simulation results and show that culturally-transmitted behaviors adapt to human learning biases (Kalish et al., 2007; Griffiths et al., 2008). Second, they have been used as a method for revealing what the biases of human learners are in contexts where these are unknown (Mesoudi et al., 2006a; Lewandowsky et al., 2009). In the language learning literature, they have provided a rich new avenue for understanding the role of learning biases (Galantucci, 2005; Kirby et al., 2008; Reali and Griffiths, 2009; Smith and Wonnacott, 2010; Galantucci et al., 2010; Verhoef, 2012; Xu et al., 2013; Silvey et al., 2014) and population structure and interaction among learners (Garrod et al., 2007; Fay et al., 2008, 2010; Garrod et al., 2010; Caldwell and Smith, 2012; Fay and Ellison, 2013) in the structure of human language (also see (Scott-Phillips and Kirby, 2010) for a review).

The social learning framework also takes an experimental approach to understanding why individuals copy the behaviors of others, in adult learners (e.g. Barrett and Nyhof, 2001; Mesoudi and Whiten, 2004; Schotter and Sopher, 2003; Kameda and Nakanishi, 2002; Efferson et al., 2008; Mesoudi and O'Brien, 2008), children (Horner et al., 2006; McEwen et al., 2007; Flynn and Whiten, 2008; Flynn, 2008), and non-human primates and other animals (Menzel, 1973; Curio et al., 1978; Sumita et al., 1985; Laland and Plotkin, 1993; Langen, 1996; Laland and Williams, 1997; Cloutier et al., 2002; Gajdon et al., 2004; Horner et al., 2006) (and see Whiten and Mesoudi (2008) for a review). Laland (2004) divides the social learning theorists' research agenda into four main questions about the cultural transmission process: 1) *what* kind of information is transmitted? 2) *who* is this information acquired from? 3) *when* do individuals decide to or end up copying others? and 4) *how* do individuals copy: via imitation, emulation, or by mediated by linguistic communication?

For example, Mesoudi and O'Brien (2008) investigates the social learning strategy of copying the most successful individual in the group (addressing the *who* question of social learning). Individuals designed virtual arrow heads in a group and received pay-offs on the basis of their design. Participants were observed to improve designs by individual, trial-and-error learning, but also by copying the designs of individuals who received higher payoffs. Trial-and-error improvements introduced new variants into the pool of arrowhead designs (i.e. the cultural variants which could be copied). However, when participants copied successful individuals, this resulted in more uniform arrowhead designs that converged upon that of the most successful participant. In addition to experiments, social learning is also rooted in a rich body of models (Campbell, 1974; Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985; Plotkin, 1994; Mesoudi

et al., 2004; Richerson and Boyd, 2005) that formalize the principles of cultural evolution in a Darwinian framework. (The work of Boyd and Richerson (1985) will be discussed further in Section 1.6.2.)

In summary, both of these frameworks approach cultural transmission with a special focus on learning. In the iterated learning literature, this focus is cognition-centric and views the replication of culturally-transmitted behaviors as a transformative process in which individual learning biases and the interaction and negotiation among learners is a central shaping force. The social learning literature approaches behaviors as more static replicators and conceptualizes learning as a direct copying process. The interesting dynamics of cultural transmission here result from different learning rules, such as "copy the majority" or "copy the most successful". In this framework, new behavioral variants are introduced when learners innovate a solution by oneself, without recourse to socially-acquired information. In this thesis, I will be adopting a more cognition-centric view of cultural evolution and argue that learning is a transformative process that both introduces variation and copies behaviors with high fidelity. Furthermore, cognition is an integrated system which, in humans, is heavily shaped by our constant social interaction with others. In this sense, any learning that humans do will be affected, in some way, by socially-mediated information in our heads. Separating individual learning from the kind of learning that happens in a social context draws an artificial line through the complexity of cognition, which may not necessarily aid progress in understanding cultural evolution.

1.4 Evolutionary theory

In 1859, Darwin revolutionized humankind's understanding of life on Earth and what it means to be human with the publication of his greatest work, *On the Origin of Species*. This book gave birth to the idea of evolution and explained two important things: the diversity of organisms in the world and the good fit of organisms to their environment. The diversity of organisms was explained by descent with modification and the goodness of fit was explained by adaptation via natural selection. Darwin described his book as "one long argument" which embodied three principles, or necessary preconditions, for evolution by natural selection (Lewontin, 1970):

1. *Variation*. Different individuals in a population must have different morphologies, physiologies, and behaviors (i.e. phenotypes).

2. *Competition*. Different phenotypes must have different rates of survival and reproduction in different environments (i.e. fitness).

3. *Inheritance*. There must be a correlation between parents and offspring such that fitness is heritable.

It is important to remember that Darwin developed the theory of evolution by natural selection without knowledge of the mechanism of inheritance. Genetic inheritance was first discovered by Mendel (1866), but did not enter scientific knowledge until its independent re-discovery by three botanists: Hugo DeVries, Carl Correns and Erich von Tschermak (Rieseberg, 1997). Evolution is a general theory that can apply to any system that satisfies the three preconditions, no matter how the mechanisms behind variation, competition, and inheritance are instantiated. Lewontin (1970) states:

> The generality of the principles of natural selection means that any entities in nature that have variation, reproduction, and heritability may evolve. ...the principles can be applied equally to genes, organisms, populations, species, and at opposite ends of the scale, prebiotic molecules and ecosystems. (p. 1-2)

Expanding further upon the types of replicators that can be subject to evolution, Szathmary and Maynard Smith (2004) describe eight "major evolutionary transitions" since the origin of life on Earth, where each one resulted in a higher-level replicator, subject to the forces of evolution. They state that primate and human culture represent the latest major transition to date. Therefore, human

cultural transmission is a good candidate system that may be subject to the forces of evolution. In his book, "Cultural Evolution", Mesoudi (2011) walks through Darwin's three preconditions and shows that human culture contains variation, competition, and inheritance, and advocates the use of general evolutionary framework for the explanation of cultural change over time. Even at the advent of evolutionary theory, its extension to cultural change and language was recognized. Darwin (1874) states "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. ...there is a limit to the powers of the memory, single words, like whole language, gradually become extinct. ... The survival or preservation of certain favored words in the struggle for existence is natural selection." Darwin (1874) also quotes Müller (1870, p. 257): "A struggle of life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand, and they owe their success to their own inherent virtue."

Since the discovery of the genetic mechanism of inheritance, evolutionary theory quickly developed and the bulk of the conceptual refinements to the theory were specialized to the way evolution operates that the molecular level. In the 1930's, the modern evolutionary synthesis took place, in which the new field of population genetics bridges micro- and macro- evolutionary theory through a variety of rigorous, formal models and mathematical proofs.

> One of the central principles of the evolutionary synthesis of the 1930s was that large-scale macroevolutionary patterns of change are the result of small-scale microevolutionary changes in gene frequencies within populations. (Mayr, 1982).

Researchers in cultural evolution are currently in a similar position to the molecular biologists and paleontologists of the 1930s. Some researchers in cultural evolution focus on the cognitive underpinnings and neural substrates that shape culture, loosely similar to the molecular level of biology, and other researchers focus on the large-scale trends. I will touch on this topic throughout this chapter and present a more thorough description of the equivalences between different fields in biological and cultural evolution in Section 1.8. Then in Chapter 6, I suggest that the problem of cultural micro- and macro-evolution synthesis is currently formulated (in the language evolution literature) as the problem of linkage (Kirby, 1999) and I provide some concrete examples of the micro-macro bridge with the data sets collected in the upcoming experiments.

Any attempt to bridge these two levels and refine the general theory of evolution for use as an explanatory tool in cultural evolution research requires a good

way of conceptualizing evolution as a general theory, which is equally applicable to molecular and cultural evolution. One such framework is provided by Sober (1984), where he describes evolutionary theory as a theory of *forces*, such as drift, selection, and mutation. A theory of forces, he says, must begin by describing the zero-force state, which characterizes the system when no forces are at play. The next step is to describe how each force acts in isolation, and then one can proceed to describe how forces act in pairs, triplets, and so on. However, it is not a priori obvious what constitutes a zero-force state for any given system and the theoretical matter of what constitutes *change* and *no change* is up to the researchers in a particular field to decide. Lloyd (1968) gives an example from two different paradigms in physics: in Newtonian theory, no change means no change in velocity, as objects remain at rest or in uniform motion when no forces impinge, but in Aristotelian theory no change means no change in location, and objects would remain at the center of the Earth when no forces impinge. As for the biological case regarding allele frequencies in a diploid population, the zero-force state is the Hardy-Weinberg equilibrium, where the proportion of the three possible diploid genotypes (AA, aa, Aa) do not change over time, despite the complicated processes involved in recombination during sexual reproduction in diploid organisms. A good question to ask ourselves here is, in the case of cultural evolution, what is the appropriate concept of a zero-force state? What are all of the conditions that would lead a language, for example, *not* to change? And what are all of the reasons why languages do change? It is up to researchers within a specific field to discover, for their own subject matter, what forces act and how they interact.

Now, I will move on to a brief introduction of the two prominent evolutionary forces: drift and selection. I will walk through these forces' formulation in the standard Wright-Fisher of population genetics and give examples of how the concept of each of these forces has been operationalized for cultural evolution.

1.5 Neutral evolution

Neutral evolution is the force that changes the frequencies of variants in a population over time due to iterated sampling error. All that is needed to study the force of neutral evolution is to relax the assumption of infinite population sizes and the frequencies of variants in a population will drift. Because all real populations are finite, neutral evolution will always be at play, to some extent.

1.5.1 Genetic drift

The first model of neutral evolution was independently proposed by two of the pioneers in population genetics, Sewall Wright (1931) and R.A. Fisher (1930). This model, now called the Wright-Fisher model of genetic drift, is the most basic model of neutral evolution and is widely applied in evolutionary biology, albeit in a great many versions with several modifications. However, all Wright-Fisher models are characterized by 1) discrete (non-overlapping) generations, 2) a fixed population size, and 3) random mating between individuals in the population (Jobling et al., 2013, section 5.3). The main concept motivating the formulation of genetic drift is finite population size:

> There remains one factor of the greatest importance in understanding the evolution of a Mendelian system. This is the size of the population. The constancy of gene frequencies in the absence of selection, mutation or migration cannot for example be expected to be absolute in populations of limited size. Merely by chance one or the other of the allelomorphs may be expected to increase its frequency in a given generation and in time the proportions may drift a long way from the original values. (Wright, 1931, p.106)

To introduce the Wright-Fisher model, let's work through an example of random mating in a finite population of hypothetical tree frogs. Suppose there exists a species of tree frogs that reproduce seasonally and have the life span of one year (satisfying the discrete generations assumption) and that each generation is composed of 10 individuals (satisfying the fixed population size assumption). Suppose also that these hypothetical frogs are haploid, with each individual carrying one of two possible alleles that determine their skin color: blue or orange. What we're interested in is how the relative frequencies of the two alleles will change over time, due only to the evolutionary force of drift. Figure 1.1 illustrates how drift can change this population in one timestep. Panel a) shows an example parent population of blue and orange alleles, where x is the frequency of the blue alleles and $N - x$ is the frequency of the orange alleles. If these individuals reproduce randomly, the population of alleles in the next generation t will be a random sample of alleles from the previous generation $(t - 1)$. Panel b) shows one possible random sample from the parent generation. The outcome of this random sample happens to be 6 blues and 4 oranges. In this fixed population of 10 frogs, there are 11 possible outcomes in generation (t), given by the following ratios of blue to orange frogs: 0:10, 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1, and 10:0. Each of these 11 outcomes is associated with a particular probability of

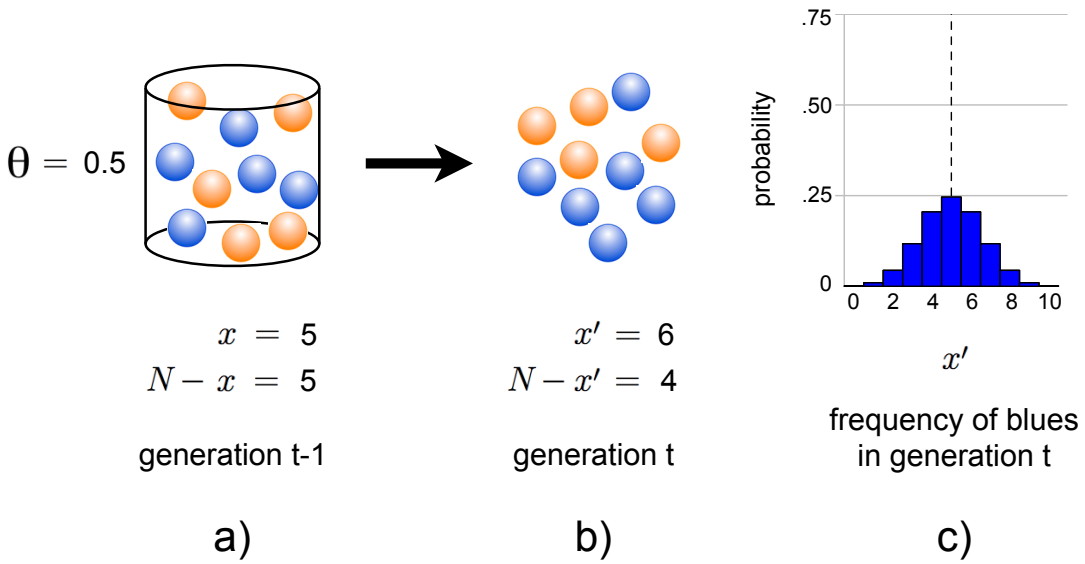


Figure 1.1: The mathematics of random sampling follows a binomial distribution, for a population of two alleles. a) The parent population of orange and blue alleles. b) A random sample of 10 alleles from the parent population. c) A binomial distribution showing the probability of getting x' blue alleles in the next generation, given a 50/50 mix of blue and orange alleles in the previous generation and a sample size of 10. The probability of getting the random sample shown in b) is 0.21.

being randomly drawn from the parent population. Panel c) gives this probability distribution over each possible outcome, plotted in terms of x' (the frequency of the blue allele in generation t). The dashed line shows the frequency of blue alleles in generation $t-1$. If we re-ran this experiment an infinite number of times, re-sampling generation t each time, we would expect to get a population of 10 blue frogs and no orange frogs 0.1% of the time, 9 blue and 1 orange 1% of the time, 8 blue and 2 orange 4% of the time, 7 blue and 3 orange 12% of the time, and so on.

The mathematical relationship between the frequency of variants (x and $N-x$) in generation $t-1$ and t is given by the binomial equation (Moran, 1958)[2]:

$$P(x'|\theta, N) = \binom{N}{x'} \theta^{x'} (1-\theta)^{N-x'} \tag{1.1}$$

This equation returns the probability of observing x' draws of the blue allele given a population of size N and a certain proportion θ. Looking back at Figure 1.1, we see that θ is the proportion of blue alleles in generation $t-1$, so $\theta = \frac{x}{N}$.

[2]Neither Wright nor Fisher express their model in terms of a binomial equation. Moran (1958) is the earliest paper I have seen to do so when describing the Wright-Fisher model.

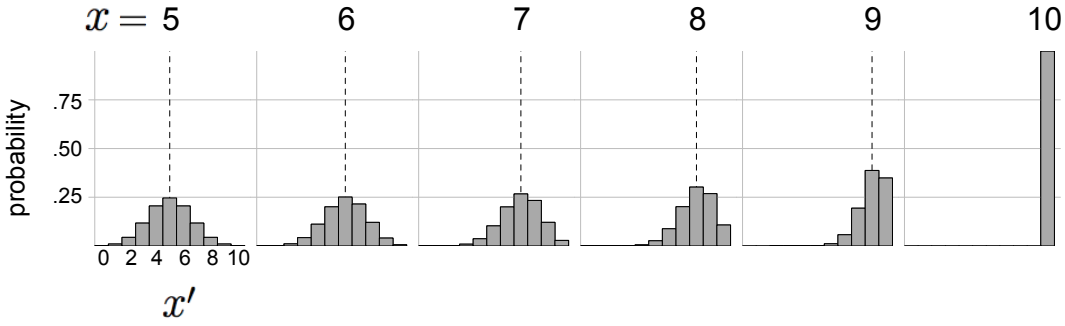


Figure 1.2: The binomial distributions that define drift for a finite population size of 10. x is the frequency of *variant* x in generation $t-1$ and x' is the frequency of *variant* x in generation t. Bar heights show the probability of observing each outcome frequency x' under drift.

Likewise, $1-\theta$ is the proportion of orange alleles in generation $t-1$. We can use Equation 1.1 to calculate the probability of obtaining the outcome shown in Figure 1.1b from the parent population in Figure 1.1a. Here, $x'=6$, $\theta=\frac{x}{N}=0.5$, and $N=10$. Plugging in these values yields $P(x'|\theta,N)=0.21$. The probability distribution in Figure 1.1c is obtained by solving this equation for each value of x'.

Figure 1.2 shows some more binomial distributions for different values of x in a population of size $N=10$. Let's abstract away from the color now and say that x could be the blue or the orange allele. Now, x is just the count of whichever allele we decide to track in the population. Each panel in Figure 1.2 shows the binomial distribution when $x=5$, 6, 7, 8, 9, and 10. (The distributions for $x=4$, 3, 2, 1 and 0 are not shown because they are just mirror images of the distributions for $x=6$, 7, 8, 9, and 10, respectively.)

A couple of things stand out here. First, the most probable outcome, regardless of the relative frequency of alleles in generation $t-1$, is no change: a population of 7 blue and 3 orange frogs is most likely to parent 7 blue and 3 orange frogs. So, drift is a process that outputs the same input frequency, with error defined by binomial variance[3]. Second, there is an absorbing state (a point of no return): when $x=10$ there is zero probability of moving to any other value of x' besides 10. Also, all values of x can lead to $x'=10$ and when this transition occurs, that means one allele is lost from the population. In the absence of a process that creates variation, such as mutation, the lost allele can never be reintroduced to the population (Ewens, 2004).

[3]The variance of a binomial distribution $= N\theta(1-\theta)$. The mean, mode, and median all coincide when $N\theta$ is an integer (Kaas and Buhrman, 1980), as will be the case in all of the examples and data sets presented in this thesis.

Equation 1.1 describes the mathematical reality of random sampling error in one generation of our frogs. But when each sample becomes the parent generation to a new generation, then this sampling error becomes iterated, and causes the frequencies of alleles to drift in the population over time. This creates an evolutionary *trajectory* through the space of all possible frequencies a variant can take in a population. Each trajectory is a timecourse sequence of transitions from $x \rightarrow x'$. All possible trajectories can, in fact, be achieved by drift (or any other evolutionary force for that matter), but it is the *likelihood* of each of these trajectories that differ under different forces of evolution. The most likely trajectory under drift is the one where our blue frogs stay at their initial frequency in the population, which is 5. However, all drifting populations will *eventually* end up in an absorbing state. In our example of coloration alleles, there are two absorbing states: all blue or all orange frogs. When a trajectory reaches an absorbing state, this is known as *fixation* (Ewens, 2004). If the population ends up with all blue frogs, then we can say that the blue allele has fixed in the population. A population can leave this absorbing state only through the introduction of new variation, via mutation or migration, for example. The fixation probability of any variant is proportional to its initial frequency in the population (Neal, 2004, p. 120). For example, if *variant x* has an initial proportion $\theta = 0.9$ and *variant y* has an initial proportion $\theta = 0.1$, the probability that *variant x* becomes fixed is nine times higher than *variant y* becoming fixed.

The rate at which a population goes to fixation depends on its size. Smaller populations have a faster average fixation rate than larger populations and will lose variation faster. The average generation at which fixation will occur ($\hat{g}$) is given by Equation 1.2 as a function of population size (N) and the proportion of *variant x* (p) (Ewens, 2004).[4]

$$\hat{g} = -2N\Big(p\log(p) + (1-p)\log(1-p)\Big) \tag{1.2}$$

For example, the average generation at which fixation will occur if $p = 0.5$ and $N = 10$ is 13.86 generations. For a smaller population ($N = 5$) fixation would occur faster at 6.93 generations, and for a larger population ($N = 100$) it would occur slower at 138.6 generations. The probability of fixation at any given generation can be calculated numerically by simulating several drift trajectories, according to the probabilities of going from $x \rightarrow x'$. Figure 1.3 shows this proba-

[4]In the literature I review, there are two notations for the proportion of *variant x* in the population. These are p and θ. These terms will always have identical meanings when used in this thesis: one can always be substituted for the other.

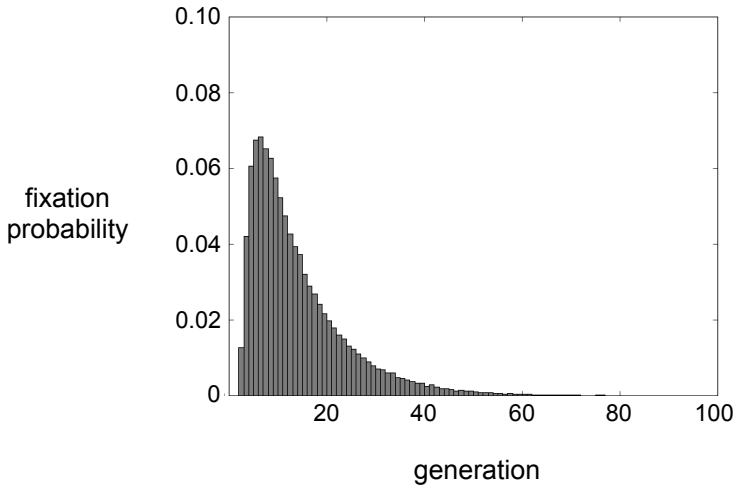


Figure 1.3: Probability of fixation per generation for drift where $N = 10$ and the initial proportion of *variant* x is $p = 0.5$. Mean fixation time is 13.86 generations.

bility distribution for population size $N = 10$, with an initial frequency of *variant* x at $p = 0.5$, for binomial drift as specified by the probabilities in Figure 1.2. 100,000 trajectories were simulated, each with an initial frequency of $p = 0.5$. Figure 1.3 plots the proportion of trajectories that fixed at each generation. This proportion is an estimate of the veridical fixation probability per generation. The mean of this distribution is 13.7, and is close to the analytical solution $\hat{g} = 13.86$.

1.5.2 Cultural drift

In the previous section, we saw how neutral change in allele frequencies occurs over time due to sampling error. Although Wright and Fisher developed their models of drift to describe change in allelic frequencies, this model is sufficiently broad to describe neutral evolution in any system where variants are sampled from old populations to create new populations. In the genetic case, actual physical entities in the world are sampled: a gene (however we decide to draw its boundaries) is physically copied from a parent to a child. In the case of unicellular organisms that replicate by fission, cells are formed by budding off of other cells. In these cases, it is intuitive that the mathematics of random sampling should apply. In the case of culture, however, cultural artifacts in the world all lose their discreteness when they enter the cognitive phase of their life cycle and it is less obvious here that the mathematics of random sampling (from physical pools of variants with replacement) will always describe neutral change in culture. Therefore, I propose that the specifics of sampling error in cultural drift is an open, empirical question, which will be taken up in detail in section

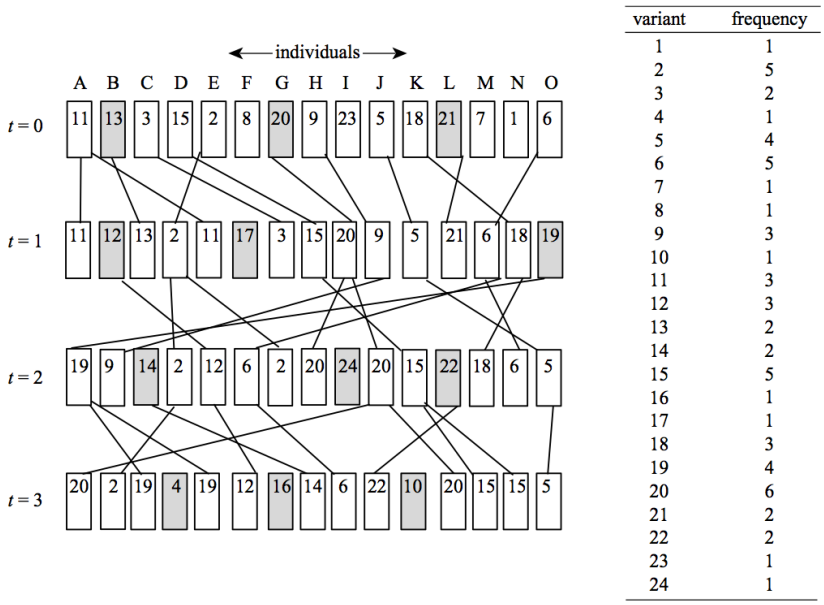
2.1, Experiment 1. The most comprehensive volume on cultural evolution to date, Boyd and Richerson (1985), only dealt with deterministic models, relegating their comments on drift to half of p.69:

> When cultural transmission is linear, the individuals who make up the next generation can be thought of as a random sample of size N of the previous generation. Thus the frequency of [variant] c after transmission, p', is a random variable with mean p and variance $(1/N)p(1-p)$. This means that if we started out with a large number of such populations, in some of them p' would be larger than p, and in others it would be smaller, but the average p' of all the population would equal p. This process we will call "cultural drift" (because it is closely analogous to genetic drift).

Here, they define drift as conforming to the binomial sampling error set forth by the Wright-Fisher model of genetic drift, in terms of binomial mean, p, and binomial variance, $(1/N)p(1-p)$. Neutral evolution, by definition should be a random sampling process that does not lead to a bias in favor of one variant over another. Therefore, as Boyd and Richerson state, the average p' of many drifting populations will equal p. If p' were consistently higher or lower than p, this would be evidence of a sampling bias and thus, non-neutral evolution. However, any kind of variance could occur and leave the necessary condition for neutral drift, $\bar{p'} = p$, unaltered.

Two other contemporaries of Boyd and Richerson, Cavalli-Sforza and Feldman (1973, 1981) and Lumsden and Wilson (1981), do incorporate drift into their models and also assume cultural drift to be binomial in nature. More recent work by Neiman (1995) and Hahn and Bentley (2003); Bentley and Shennan (2003); Herzog et al. (2004); Bentley et al. (2004); Bentley (2008) have applied the binomial drift model to cultural data sets and shown that a variety of cultural change conforms to this type of drift. Bentley and colleagues investigated a wide variety of data sets: first names, pottery motifs, patent citations, lexical items in academic publications, and dog owners' choice of breed. They analyze data at a macro level, plotting the frequency of cultural variants against their rank, and assess whether or not this distribution conforms to a power law. Random sampling processes produce power law distributions when data is formatted in this way. Bentley et al. (2004) demonstrate this fact algorithmically, by simulating the random sampling of cultural variants from a population (Figure 1.4).

The left-hand side of Figure 1.4 gives a schematic representation of their random copying algorithm. Each row is a population of people (A through O) in generation t and each person can have one cultural variant (of type 1 through



variant	frequency
1	1
2	5
3	2
4	1
5	4
6	5
7	1
8	1
9	3
10	1
11	3
12	3
13	2
14	2
15	5
16	1
17	1
18	3
19	4
20	6
21	2
22	2
23	1
24	1

Figure 1.4: Diagram of Bentley et al. (2004)'s simulation of the random copying of cultural variants. Reproduced from Bentley et al. (2004, p. 1444).

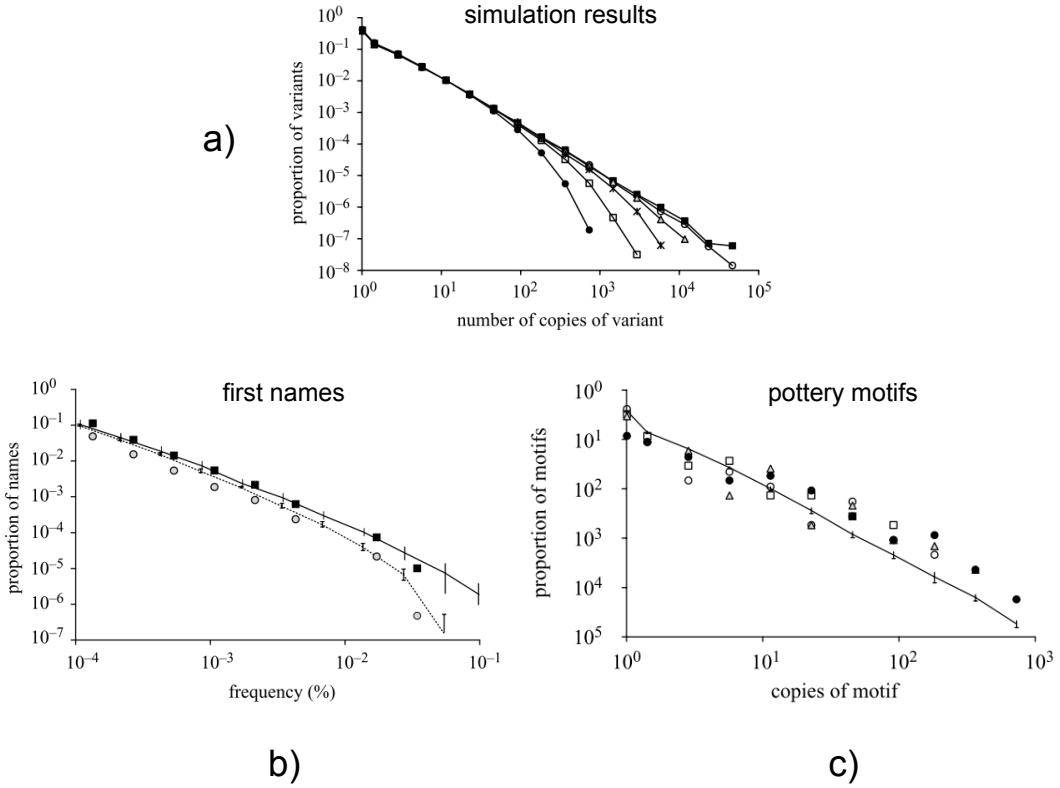


Figure 1.5: a) The power law distributions resulting from the simulation in Figure 1.4 for 6 values of $\mu = 0.128, 0.064, 0.032, 0.016, 0.008, 0.004$ (going left to right across the tails), 1000 generations, and population size of 250. Each line is the average of 5 separate runs. b) US census first name data. b) Neolithic pottery motif data. Reproduced from Bentley et al. (2004, p. 1445, 1446, & 1447).

infinity). Subsequent generations are formed by randomly sampling a variant from the previous generation (each copying event is shown by a connecting line) or innovating a new variant with probability μ, which for this example is 3 mutations per generation ($\mu = 3/15 = 0.2$). The simulation is run for many generations and a cumulative tally of each variant's frequency is logged. The logged tallies for the four generations in the schema are shown to the right. The results of this simulation are shown in Figure 1.5a. The x-axis shows the number of copies of each variant on a log scale and the y-axis shows the number of variants that had each frequency. Most variants in the simulation (about $10^{-0.5} = 5\%$) only existed in one copy (10^0). And only a couple of variants are highly frequent (in the 10^3 to 10^5 range). Six lines are plotted for different mutation rates and higher mutation rates lead to distributions with tails that drop off more in the lower frequency ranges. This is because the mutations convert many would-be copies of the high-frequency variant into new variants that have a frequency of 1, culling the numbers of higher-frequency variants. Figure 1.5b and c show two example data sets which are well-fit by straight lines on a log-log plot, indicating a power law distribution (cf. Clauset et al., 2009).

Although Bentley interprets these results as examples of neutral cultural evolution, plausible alternative hypotheses should be tested and ruled out to strengthen this claim. However, the existing literature provides few clear alternatives in which non-neutral copying mechanisms produce power-law distributions, so this is an open area for investigation. A few positive examples exist in the natural language processing (NLP) literature (Goldwater et al., 2005, 2011) and the iterated learning literature (Tullo and Hurford, 2003). (Goldwater et al., 2005, 2011) develop two models in response to the failure of standard statistical NLP models to capture the power-law distributions of word token frequencies, which is a robust property of natural languages (Estoup, 1916; Zipf, 1935, 1949). These are computational-level stochastic models of cognitive processes that entail non-neutral, frequency-dependent copying of variants. Additionally, Tullo and Hurford (2003) developed a "discourse-triggered meaning choice" model in which interacting, communicating agents make frequency-dependent choices of words from past conversational topics. Mesoudi and Lycett (2009) also re-implement Bentley's model with two frequency-dependent copying strategies (at a non-cognitive, social learning level), however these fail to produce power law distributions and will be further described in Section 1.6.2. They do state, however, that it is theoretically possible for opposing forces of non-neutral copying to cancel one another out and produce macro-level evolution that appears identical to drift at the macro level.

Before leaving this section, I would like to mention, for the purpose of disambiguation, that the term *drift* as used in the historical linguistics and sociolinguistics literature does not refer to a neutral copying process, but "stochastic directional change", and was borrowed into these fields from the concept of "continental drift", coined by Wegener (1912). Sapir (1921) refers to many *drifts* in historical language change, offering explanations for all of them. One example is the drift toward invariable word order (p. 186), which he explains is due to the gradual loss of case marking across the world's languages. These are clearly directional, biased sampling processes. However, when drift exclusively refers to neutral evolution, it can only have one explanation: sampling error.

A passage from Vennemann (1975), entitled "An explanation of drift" provides a clear characterization of the historical/sociolinguistic usage of drift:

> If all individual deviation from a linguistic norm were of equal status, equally probable to be accepted by the speech community and to become a new norm, they would cancel each other out so that a linguistic norm would not change or would only vacillate insignificantly in a way that reflects the variation among individual speakers. Since linguistic norms neither remain unchanged nor change back and forth but change in certain directions, and fairly rapidly so, it cannot be the case that individual deviation from the norm have equal status. Rather, certain deviation, or certain types of deviations, must be favored over others, must be more readily produced and must definitely be more readily accepted. Since such cumulative favorization occurs, at least for the most part, without conscious awareness on the side of the language users, it must be rooted in general psychological tendencies. If we can identify these tendencies, we can predict the future course of a language, the direction of its "drift". We can "prophesy". (p. 271)

Drift is an attractive term for historical and sociolinguists who want to emphasize the stochastic aspects of evolutionary change and move away from deterministic explanations for the drivers of historical language change. However, I think that researchers in cultural evolution should strive for a consistent terminology and maintain a concept of drift that purely refers to neutral evolution. This provides an appropriate baseline, or null hypothesis for cultural change, that we can use as a tool for identifying the intricate dynamics of all forms of directional and biased sampling that occur in culture for diverse reasons.

1.6 Selective evolution

In this section we take up a force behind evolutionary change that can create adaptation of organisms to their environment. This is the force of selection, which constitutes a form of biased sampling error that can lead to directional change over time.

1.6.1 Genetic selection

Wright (1931) and Fisher (1930) also described how selection would act in finite populations. Selection is a process that favors or disfavors particular variants on the basis of some criteria, such as coloration, causing differential survival or reproduction among variants. We can talk about selection as being positive or negative: *positive selection* favors a particular variant and *negative selection* disfavors it. In the example of our hypothetical tree frogs, a selective pressure such as a predator that is better at picking out the orange frogs against the jungle foliage, will tend to increase the relative proportion of blue to orange frogs in every generation. This selection pressure can be formulated as negative selection against the bright color of orange frogs, or as positive selection for the good camouflage of blue frogs.

We can also talk about selection in terms of the absolute and relative fitness of variants. Absolute fitness is quantified by the average number of offspring an organism with the favored allele leaves and is equivalent to the multiplication rate of the variant in question. Say for example, our population of frogs is at 40 blues and 60 oranges in generation 1, and then 80 blues and 90 oranges in generation 2. Here, the absolute fitness of the blue frogs is 2 (because each frog leaves an average of 2 progeny) and the absolute fitness of the orange frogs is 1.5. Because natural selection is about the differential reproduction and survival of variants, most population genetics models deal with measures of relative fitness, rather than absolute fitness (Neal, 2004, p. 147). To translate the absolute fitness measure into a relative fitness measure, one variant must be chosen as the reference variant. Conventionally, the variant with the higher growth rate is designated as the reference variant and thus, all other variants are discussed in terms of negative selection (Neal, 2004, p. 147). Table 1.1 lays out the relationship between absolute fitness, relative fitness, and the selection coefficient, in terms of positive selection for the blue frogs (left) and negative selection against the orange frogs (right). After determining the average growth rate of each variant (2 for blue frogs and 1.5 for orange frogs), relative fitness is derived by dividing the

	positive selection on blue (orange is reference)			*negative selection on orange (blue is reference)*		
variant	blue	orange	total	blue	orange	total
initial frequency	40	60	100	40	60	100
next generation frequency	80	90	170	80	90	170
absolute fitness (λ)	2.0	1.5		2.0	1.5	
relative fitness ($W = \frac{\lambda}{\lambda_{ref}}$)	$\frac{2.0}{1.5} = 1\frac{1}{3}$	$\frac{1.5}{1.5} = 1$		$\frac{2.0}{2.0} = 1$	$\frac{1.5}{2.0} = \frac{3}{4}$	
selection coefficient ($s = 1-W$)	$1-1\frac{1}{3} = \frac{1}{3}$	$1-1 = 0$		$1-1 = 0$	$1-\frac{3}{4} = \frac{1}{4}$	

Table 1.1: Example calculations showing the relationship between growth rate, absolute fitness, relative fitness, and the selection coefficient. The choice of reference variant determines whether or not the selection coefficient represents positive or negative selection. When working with negative selection coefficients, s must be made negative before plugging it into Equation 1.3.

growth rate of the variant in question by the growth rate of the reference variant. The selection coefficient is simply 1 minus the relative fitness ($s = 1 - W$).

The value of the selection coefficient will differ whether we are talking about positive selection on one variant or negative selection on the other. In Table 1.1, the negative selection coefficient on the orange frogs tells us that the orange frogs are one fourth worse off than the blue frogs, whereas the positive selection coefficient on the blue frogs tells us that the blue frogs are one third better off than the orange frogs. Negative selection coefficients are bounded by 0 (no selection) and -1 (complete lethality). Positive selection coefficients are bounded by 0 (no selection) and infinity, because growth rates do not have an upper bound in theory[5]. Because the -1 to 0 bounds are easier to work with, the negative representation of the selection coefficient is more commonly used.

These measures of fitness and selection are all deterministic: an infinitely large population would be composed of exactly one third more blue frogs than orange frogs. To understand how selection might operate in a finite population, the selection coefficient can be added into the Wright Fisher model as a weight on θ in the binomial sampling equation (Equation 1.1). Here, the selection coefficient is proportionalized into $\frac{x(1+s)}{x(1+s)+(N-x)}$ and plugged into θ:

$$P(x'|\theta, N) = \binom{N}{x'} \left(\frac{x(1+s)}{x(1+s)+(N-x)}\right)^{x'} \left(1 - \left(\frac{x(1+s)}{x(1+s)+(N-x)}\right)\right)^{N-x'} \tag{1.3}$$

[5]whereas death rates do have an upper bound, in theory *and* in practice. (All further footnotes will be less ominous than this one...)

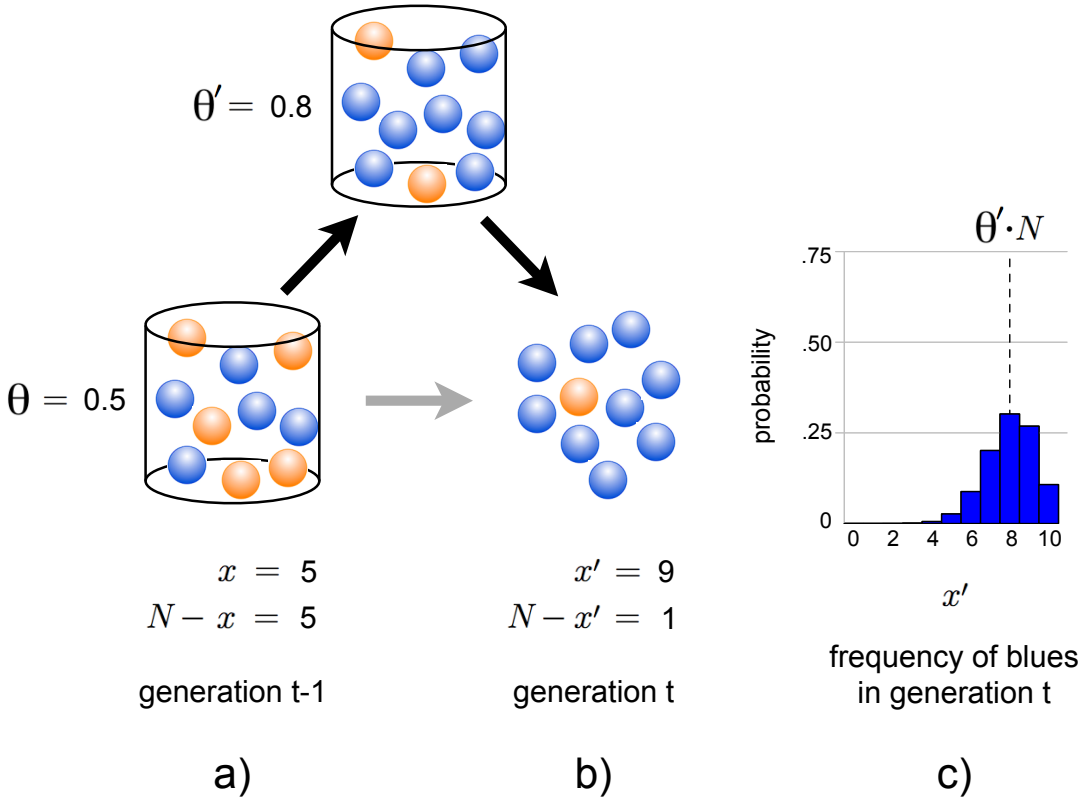


Figure 1.6: Biased sampling error due to selection in a finite population of two alleles. a) The parent population of orange and blue alleles. b) A random sample of 10 alleles from the parent population, mediated by selection (represented above center as an alteration to the probability that each allele type will be sampled). c) A binomial distribution showing the probability of getting x' blue alleles in the next generation due to selection and sampling error for a population of size 10. The probability of getting the sample shown in b) is 0.27. The mean of this distribution equals $\theta' \cdot N$.

Given its formulation in Equation 1.3, it is clear that the selection coefficient alters the probability of sampling variant x from the veridical proportion of variants in generation $t-1$. This sampling process will be biased in a particular direction (depending on whether s is positive or negative) and with a particular strength (depending on the magnitude of s). One way to think about this process is as if population t were not sampled from the veridical $t-1$ population, but instead from some imaginary $t-1$ population, where the frequency of *variant* x is adjusted by s. This conceptual framing is depicted in Figure 1.6. Panel a) shows the parent population of blue and orange alleles, where x is the frequency of the blue alleles and $N-x$ is the frequency of the orange alleles, and panel b) shows a random sample from the parent population, mediated by a selection pressure that favors blue alleles. In this example, a selection coefficient of $s = 0.65$ alters $\theta = 0.5$

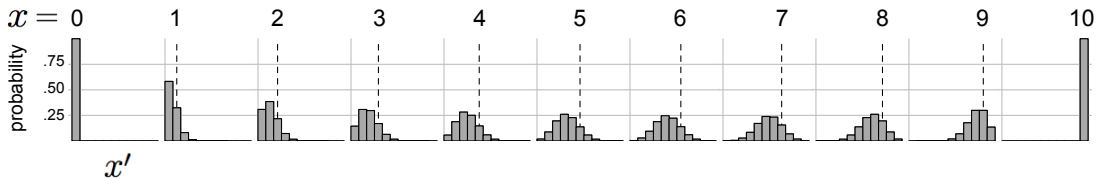


Figure 1.7: Wright-Fisher with negative selection ($s = -0.5$) against *variant x*. x is the frequency of *variant* x in generation $t-1$ and x' is the frequency of *variant* x in generation t. Each y-axis ranges from $x' = 0$ to 10. Bar heights show the probability of observing each outcome frequency x' under this type of selection.

to a new value, $\theta' = 0.8$. Each of the 11 possible outcomes in generation t has a particular probability of being randomly sampled from θ'. Panel c) gives the probability distribution over each possible outcome, plotted in terms of x' (the frequency of the blue allele in generation t). If we re-ran this experiment an infinite number of times, re-sampling generation t each time, we would expect to get a population of 10 blue frogs and no orange frogs 11% of the time, 9 blue and 1 orange 27% of the time, 8 blue and 2 orange 30% of the time, 7 blue and 3 orange 20% of the time, and so on. The dashed line shows the mean of this distribution, which is dictated by θ'. Here, $\theta' \cdot N$ is the *expected frequency* of *variant* x in the next generation and θ' is the *expected proportion*.

Figure 1.7 shows the distributions of x' from the Wright Fisher model with selection, for negative selection of $s = -0.5$ against *variant* x, and a population size of 10. Each distribution shows the probability of each outcome of x' given one value of x. Unlike drift (refer back to Figure 1.2), the expected frequency of *variant* x is not equal to its frequency in the previous generation. Instead, it is shifted toward the variant that is being selected for (*variant* y in this case). In the middle panel, for example, the input frequency of *variant* x was 5 and the most likely frequency of *variant* x in the next generation is $x' = 3$. This shift in expected frequency is due to the sampling bias that selection imposes on the population each generation.

Selection defines a mapping between each veridical proportion (θ) and the expected proportion (θ') and this mapping is known as a selection function. Figure 1.8 plots the selection functions for some example selection coefficient values. Panel a) plots the selection function for the coefficient used in Figure 1.7 where $s = -0.5$. Here we see more clearly how the negative selection against *variant* x lowers its expected proportion (θ') relative to its proportion in the previous generation (θ). For example, when the proportion in the previous generation is 0.5, then the expected value in the next generation will be 0.3. Panel b) plots a

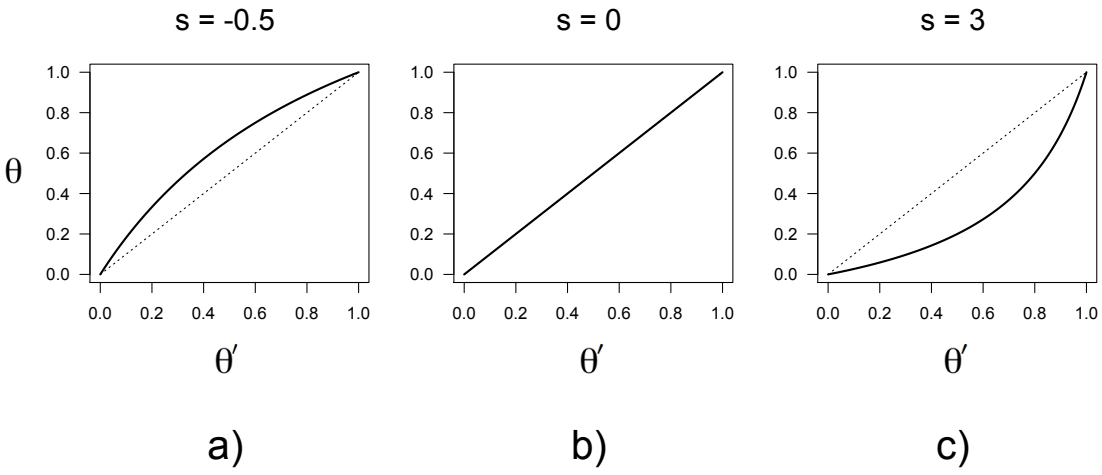


Figure 1.8: Some example selection functions, which define the relationship between the expected proportion of *variant* x (θ') and its proportion in the previous generation (θ). The dotted line marks the points where $\theta = \theta'$. a) Negative selection where $s = -0.5$. b) No selection where $s = 0$. c) Positive selection where $s = 3$.

selection coefficient of $s = 0$ where there is no selection. Here, $\theta = \theta'$ and this system is identical to that described by drift. Panel c) plots a selection coefficient of $s = 3$, which yields strong positive selection for *variant* x. For example, if *variant* x constitutes 50% of the population in generation $t - 1$, it is expected to constitute 80% of the population in generation t.

Because the plots in Figure 1.8 map these changes in terms of proportions, they apply to any population size. Increasing the population size will decrease the sampling error around the expected frequency, but it will not change the expected frequency. When the population size is infinite, there will be no sampling error. Therefore, these plots model the actual, deterministic behavior of the system if the population size were infinite. The models of cultural selection that will be presented in the next section take the form of deterministic models for populations of infinite size. However, the stochasticity of these models can be recovered by specifying the sampling error for a given population size.

The selection functions depicted in Figure 1.8 assign constant fitness values to variants on the basis of their color. However, many more selection functions can be formulated when fitness is determined by a variant's relative frequency in the population. This is known as frequency-dependent selection and is defined by any selection function that allows the selection coefficient to vary as a function of the allele's frequency. Returning to our populations of frogs, examples of frequency-dependent selection pressure might be a predatory bird that prefers to eat frogs of

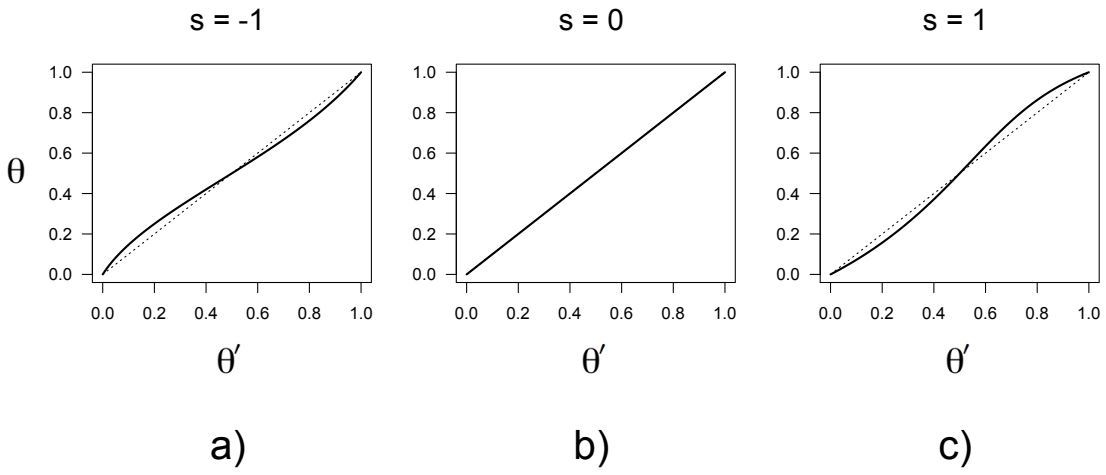


Figure 1.9: Some examples of frequency-dependent selection functions, which define the relationship between the expected proportion of *variant* x (θ') and its proportion in the previous generation (θ). The dotted line marks the points where $\theta = \theta'$. a) Frequency-dependent selection against the rare variant where $s = -1$ and $f = -0.5$. All proportions under 0.5 will be slightly under-represented in the next generation and all proportions over 0.5 will be slightly over-represented. b) The neutral case where no selection is acting: $s = 0$ and $f = 0$. c) Frequency-dependent selection that favors rare variant where $s = 1$ and $f = 0.5$. All proportions under 0.5 will be slightly over-represented in the next generation and all proportions over 0.5 will be slightly under-represented.

the rare color, or a preference among the frogs for mating with frogs of the more common color. In both of these cases, the rare allele will quickly be eliminated from the population. On the other hand, if the predatory bird prefers the more common color, or the frogs prefer to mate with the rare color, then the common allele will decrease in the population, and if it becomes rare it's frequency will start to go up again.

An example model of frequency-dependent selection for the rare allele is given by Felsenstein (2005), p.113, which I have revised in terms of θ and θ':

$$\theta' = \frac{\theta(1 + f - s\theta)}{1 + (f - s\theta)\theta} \tag{1.4}$$

The sign of the selection coefficient (s) determines whether there is positive or negative selection for the rare allele and f modifies the shape of the selection function. Figure 1.9 shows example selection functions given by Equation 1.4.

These models of selection all have absorbing states: when $x = 0$ or 10 there is zero probability of moving to any other value of x' besides 0 or 10, respectively. Both drift and selection will eventually lead to the loss of one variant, but a crucial point here is that most forms of selection will, on average, eliminate variation

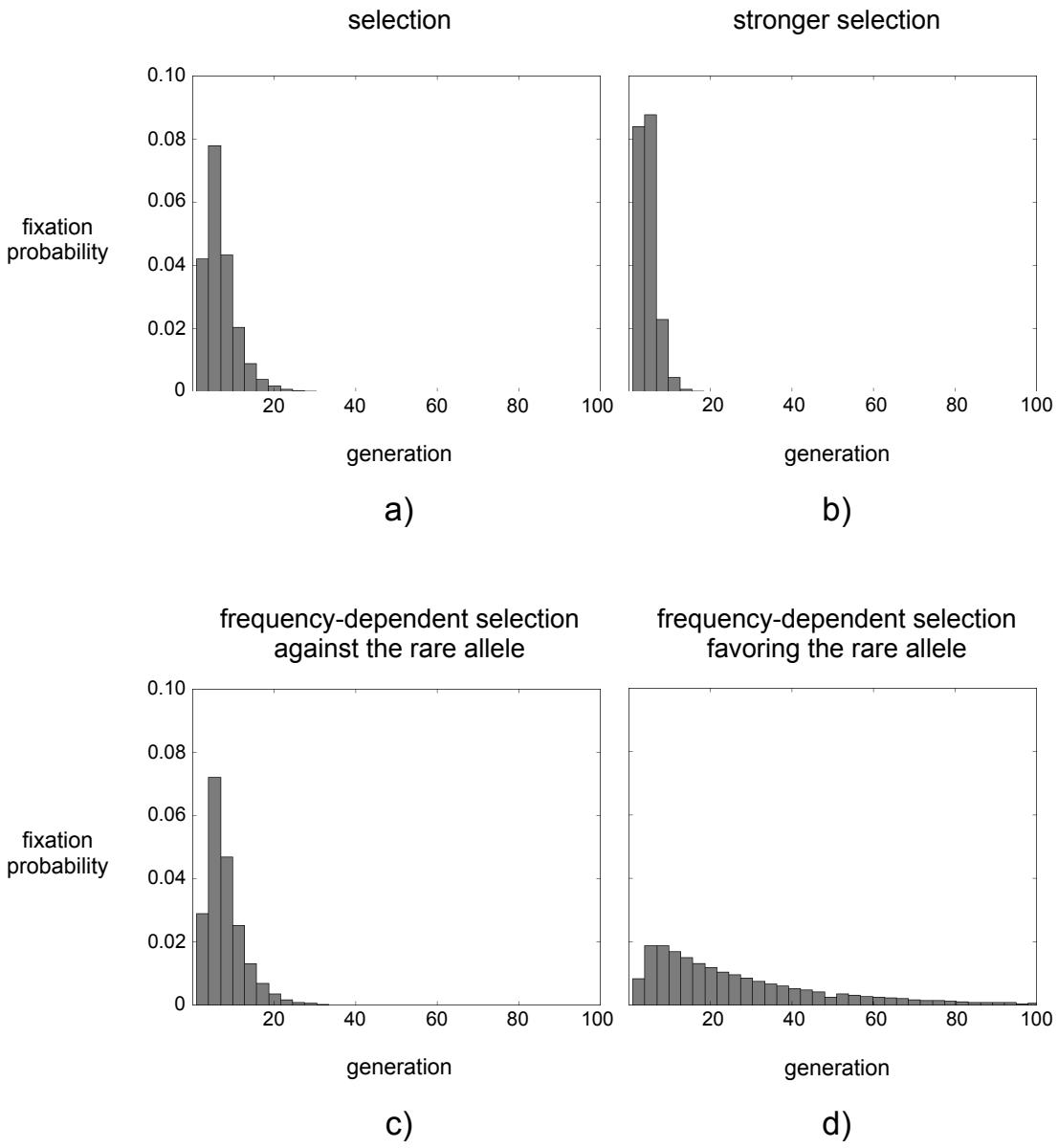


Figure 1.10: Probability of fixation per generation for four types of selection, where $N = 10$ and the initial proportion of *variant* x is $p = 0.5$. a) Selection where $s = 0.5$, mean fixation time is 6.59 generations. b) Selection where $s = 1$, mean = 4.34. c) Frequency-dependent selection against the rare allele, where $s = -1$ and $t = -0.5$ (same model as shown in Figure 1.9a). d) Frequency-dependent selection in favor of the rare allele, where $s = -1$ and $t = -0.5$ (same model as shown in Figure 1.9c).

faster than drift. Figure 1.10 shows the fixation distributions for four different models of selection. Comparing panels a) and b) we see that stronger selection leads to the faster elimination of a variant. Likewise, both of these models eliminate variation faster than expected under drift, where mean fixation time was 13.86 generations (refer back to Figure 1.3). Panels c) and d) show two types of frequency-dependent selection. In the case of selection against the rare allele, fixation also occurs faster than expected under drift, nearly always by eliminating the rare allele. However, when selection favors the rare allele, variation is maintained in the population much longer than would be expected under drift. This type of selection is one of the main mechanisms behind balancing selection, which will be discussed again in Chapter 6.

Before moving on, I would like to reiterate one point. If we step back for a more general view on these basic population genetics models, we see that there are two clear components: a deterministic one that encompasses *sampling bias* and a stochastic one encompassing the *sampling error*. By definition, neutral evolution is change due only to the stochastic component, and selective evolution is due to the presence of a sampling bias, however small or large it may be.

1.6.2 Cultural selection

Given the formulation of selection as sampling bias, any cultural transmission or copying process that over- or under-represents particular variants can be understood as a form of selection. Boyd and Richerson (1985) provide the most comprehensive catalogue, to date, of the range of forces behind cultural evolutionary change. They sketch five broad classes of "conceivable processes that could change culture through time" (p. 9):

1. *Random variation.* Errors in the neurological machinery of the mind, analogous to mutation, that provide an unbiased source of new cultural variants. This is likely to be much higher than the rate of genetic mutation.

2. *An analog of genetic drift.* Chance variations in which cultural variants are observed and remembered, which may cause substantial changes in the frequency of variants over time, especially when population sizes are small (i.e. when the number of cultural variants is small).

3. *The force of guided variation.* Individuals' adjustment of their own cultural variants via learning, such as trial-and-error learning, and rational calculation. In contrast to random variation, innovations that result from learning

and rational calculation are not randomly generated. This leads to a type of Lamarckian inheritance that increases the frequency of the cultural variants that are more likely to be generated.

4. *Biased transmission.* When the cultural transmission process itself favors some cultural variants over others. This is similar to guided variation because both forces arise from the same capacities for learning and rational calculation. However, the forces of biased transmission are more complex because they are affected by their own output (creating strong feedback cycles) more than guided variation is.

5. *Natural selection can operate on culture.* If the set of cultural variants one possesses can differentially affect their reproduction and survival, then these cultural variants can be the target of natural selection and those that positively effect survival and reproduction will increase in frequency relative to others.

A large body of work exists on the fifth class: the role of natural selection on cultural traits, as an aspect of an organism's phenotype (e.g. Cavalli-Sforza and Feldman, 1973; Wilson, 1975; Symons, 1979; Alexander, 1980; Hinton and Nowlan, 1987; Feldman and Laland, 1996; Odling-Smee et al., 2003; Chater et al., 2009; Davies et al., 2012). However, I am going to leave this class of forces behind and focus throughout the remainder of this thesis on the evolutionary forces that originate from types of sampling bias and sampling error that cognition exerts on culturally transmitted behavior (classes 1 through 4). Although cognitive architecture and processes certainly have a genetic basis, this may not be the most illuminating level of explanation for why distributions of cultural traits regarding, for example, cuisine, clothing, music, language, and technologies are the way they are and how they change over time. The data sets I will be dealing with in this thesis contain cultural evolutionary changes that happen within the lifetime of an individual and therefore, cognition is the appropriate level at which explanations of these, and perhaps most, forms of cultural evolution should be addressed (cf. Scott-Phillips et al., 2011).

So, focusing now on classes 1 through 4, Boyd and Richerson only define guided variation and biased transmission as a source of sampling bias. The other two are described as random sources of variation, akin to random mutation. However, to a cognitive scientist, the examples Boyd and Richerson give for random cultural variation (errors in the neurological machinery of the mind) and cultural drift (errors in how cultural variants are observed and remembered) define clear

sources of sampling bias. Furthermore, the distinction between their two sources of sampling bias, guided variation and biased transmission, which both result from the "same capacities for learning and calculation" (p. 10), only exists because a line has been drawn between cultural information that is acquired via individual learning and cultural information acquired via social learning.

Because only biased transmission deals exclusively with socially-acquired information, it is of primary interest to the study of cultural evolution from the social learning perspective. First, Boyd and Richerson describe cultural transmission as the copying of cultural variants by naïve individuals. If naïve individuals copied variants at random, transmission would be unbiased. However, if individuals discriminate between variants and copy them for particular reasons, then biased transmission occurs. They identify three classes of biased transmission:

1. *Direct bias.* This occurs when individuals discriminantly copy variants on the basis of some intrinsic property of the variant, because they prefer it for some reason. Individuals need not have the same preferences, but so long as individuals are copying on the basis of their own preferences, direct bias occurs.

2. *Frequency-dependent bias.* This occurs when individuals copy a variant on the basis of its frequency. Any conceivable function that maps frequencies to copying probabilities would fall under this category, although they focus on two: *conformity copying*, in which individuals copy the most frequent variant(s), and *anti-conformity copying*, in which individuals copy the least frequent variant(s).

3. *Indirect bias.* This occurs when individuals copy a variant on the basis of some factor that is correlated with the variant, such as the social prestige of the individuals who exhibit that variant.

These forms of biased transmission can be understood as forms of selection pressures on a pool of cultural variants. I will explain the first two forms in more detail. Direct bias operates similarly to selection, as described in the previous section, because it leads to an over-production of a variant on the basis of an intrinsic property. Just as our hypothetical frogs may have been selected for on the basis of their color, cultural variants can be selectively copied on the basis on an intristic property of that variant, such as the choice to wear an orange or a blue wig on Queen's Day in the Netherlands[6].

[6]or whether or not to paint your toddler in blackface on Sinterklaas Day.

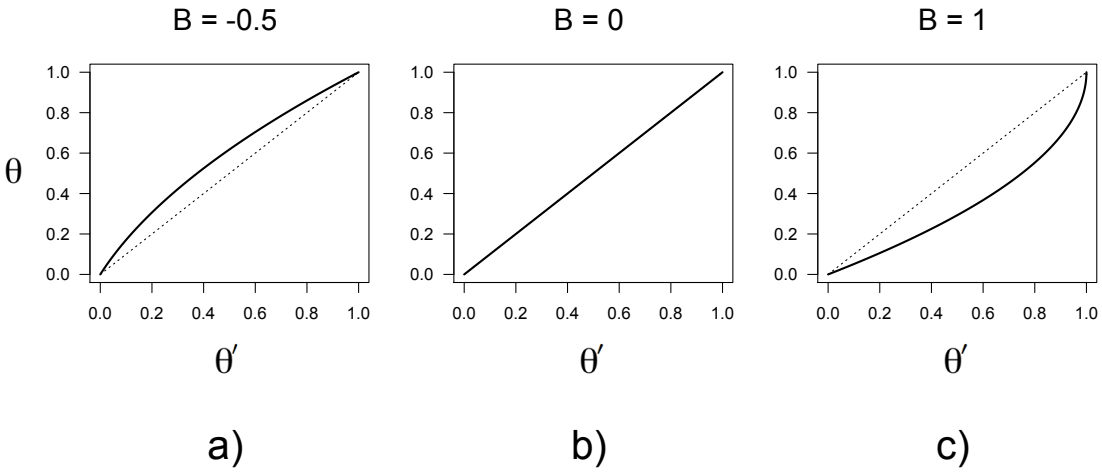


Figure 1.11: Some examples of direct bias selection functions, which define the relationship between the expected proportion of *variant x* (θ') and its proportion in the previous generation (θ). The dotted line marks the points where $\theta = \theta'$. a) Direct bias in favor of *variant y*, where $s = -0.5$. b) Unbiased copying, where $s = 0$. c) Direct bias in favor of *variant x*, where $s = 1$.

Boyd and Richerson (1985, p.138) formalize direct bias in the following equation, which I have revised in terms of θ and θ':

$$\theta' = \theta + B\theta(1 - \theta) \tag{1.5}$$

where θ is the proportion of *variant x* in generation $t - 1$, θ' is the proportion of *variant x* in generation t, and B is a parameter determining the strength and direction of the direct bias. When B is positive, *variant x* is selected for, and when B is negative, *variant x* is selected against. Figure 1.11 shows some example selection functions as determined by direct bias in Equation 1.5.

Also, frequency-dependent bias can be understood as frequency-dependent selection, because it leads to an over-production of a variant on the basis of that variant's frequency in the population. Boyd and Richerson (1985, p.208) give the following equation for frequency-dependent bias, revised in terms of θ and θ':

$$\theta' = \theta + D\theta(1 - \theta)(2 - \theta - 1) \tag{1.6}$$

where D is a parameter that determines the strength and direction of bias for the more common cultural variant. When D is negative, there is a bias in favor of the rare variant and this constitutes anti-conformity copying. When D is positive, the common variant is favored and conformity copying occurs. Figure 1.12 shows some example selection functions as determined by frequency-

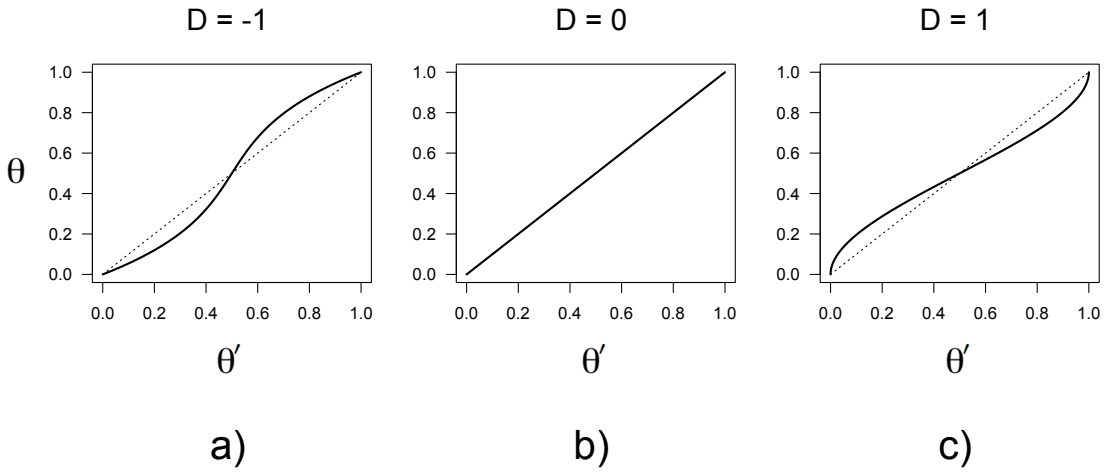


Figure 1.12: Some examples of frequency-dependent bias selection functions, which define the relationship between the expected proportion of *variant* x (θ') and its proportion in the previous generation (θ). The dotted line marks the points where $\theta = \theta'$. a) Frequency-dependent bias in favor of the rare variant where $D = -1$. All proportions under 0.5 will be slightly over-represented in the next generation and all proportions over 0.5 will be slightly under-represented. b) Unbiased copying, where $D = 0$. c) Frequency-dependent bias in favor of the common variant where $D = 1$. All proportions under 0.5 will be slightly under-represented in the next generation and all proportions over 0.5 will be slightly over-represented.

dependent bias in Equation 1.6. Although the shapes of these curves differ from the frequency-dependent selection model described in Section 1.6.1, both of these models constitute frequency-dependent selection because they selectively favor variants on the basis of their frequency in the population. Additionally, these are just two of an infinite number of frequency-dependent selection functions that mathematically exist (e.g. Lachmann-Tarkhanov and Sarkar, 1994).

Boyd and Richerson (1985) only describe the forces of cultural evolution with deterministic models (as depicted in Figures 1.11 and 1.12), but they do state that finite, culturally evolving populations would also be subject to drift, and that this force of drift would be in the form of binomial / multinomial sampling error (refer back to their quote on this in Section 1.5.2). Although they do not conduct any analyses of cultural evolution in finite populations, this form of sampling error can be added to their models by substituting θ in the Wright-Fisher model (Equation 1.1) with θ' from any model of interest.

All the forms of cultural selection discussed thus far address cultural evolution at the micro level: in terms of individuals' propensities to copy certain cultural variants on the basis of certain criteria. I will now discuss some macro-level effects of cultural selection and contrast this to the macro-level cultural drift literature described in Section 1.5.2.

Mesoudi and Lycett (2009) implement Boyd and Richerson's two frequency-dependent copying strategies (conformity and anti-conformity copying) as modifications to the cultural drift model of Bentley et al. (2004) and show how these two forces of cultural evolution differ from drift. Figure 1.13 shows my own implementation of the Bentley et al. (2004) model with conformity and anti-conformity transmission biases. Both copying strategies are implemented by raising the veridical distribution of cultural variants at each generation to a power, given by the parameter r, and sampling the next generation of variants from the resulting distribution. When $r > 1$ this implements conformist copying and results in the higher-frequency variants being copied disproportionately more. When $r < 1$ this implements anti-conformist copying, where lower-frequency variants are copied disproportionately more. These results replicate those of Mesoudi and Lycett (2009), though the implementation of the copying biases is slightly different.

In Figure 1.13, we see that conformist copying lowers the number of variants that exist with few copies compared to drift. In the cultural drift plot, there are $\approx$ 10,000 (10^4 on the y-axis) variants that exist in only one copy (10^0 on the x-axis). Whereas in the conformist copying plot there are $\approx$ 1000 variants that exist in only one copy. Also, conformist copying creates a long tail to the power law in which a handful of variants exist in extraordinarily high proportions in the population. This is a form of winner-take-all dynamic. Anti-conformity copying also lowers the number of variants that exist in one copy by an order of magnitude (for these parameter settings). Although individuals are selection low-frequency variants, low-frequency variants quickly rise in numbers and then become undesirable to copy. The population-level effect of this copying strategy is that most variants exist in a the mid-ranged number of copies.

Mesoudi and Lycett (2009) call attention to the fact that macro-level states, such as power law distributions, are multiply realizable: different individual copying strategies may give rise to the same macro-level phenomena. They present a further type of frequency-dependent copying model, in which individuals never copy rare traits, and show that this also yields a distribution that appears to be a power law. They also state that opposing selection pressures could lead to drift-like power laws as well. Although we can see from the plots in Figure

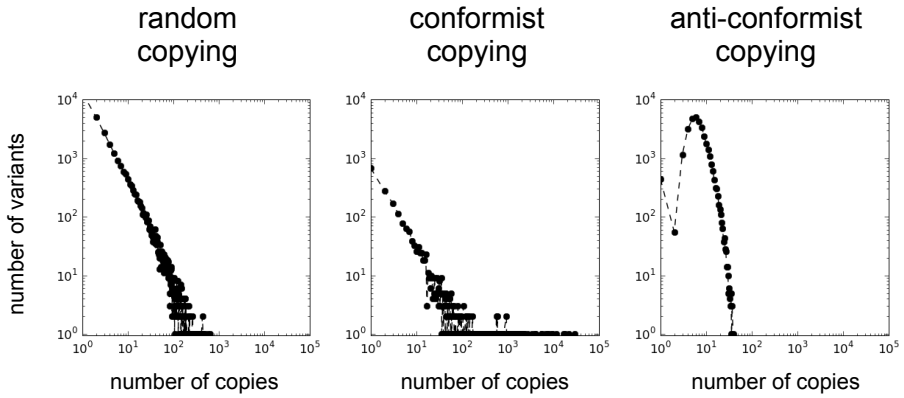


Figure 1.13: Replication of Bentley et al. (2004)'s random copying model (left). Implementation of Boyd and Richerson (1985)'s conformist copying model (middle) and anti-conformist copying model (right), which replicate Mesoudi and Lycett (2009). Each simulation uses population size 250, 1000 generations, and mutation rate $\mu = 0.128$. The conformist model uses $r = 2$ and the anti-conformist model uses $r = -2$.

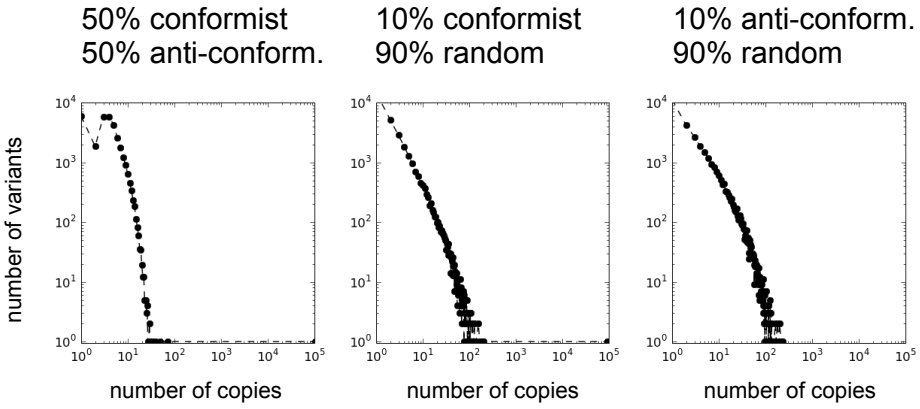


Figure 1.14: Novel implementations with populations of mixed learners. Left: 50/50 mix of conformist and anti-conformist copiers. Middle: 10% conformist copiers in a population of random copiers. Right: 10% anti-conformist copiers in a population of random copiers. Each simulation uses population size 250, 1000 generations, and mutation rate $\mu = 0.128$. The conformist model uses $r = 2$ and the anti-conformist model uses $r = -2$.

1.13 that conformity and anti-conformity copying do not appear to be opposing forces (despite their neatly juxtaposed names), I was interested to see what the additive effect of having these two copying strategies in the same population may be. Figure 1.14 shows an implementation of mixed populations of conformist and anti-conformist copiers (left), random and conformist copiers (middle), and random and anti-conformist copiers (right). In the left plot, the aspects of each copying strategy are clearly visible: conformity copiers contribute to the long tail and anti-conformity copiers contribute to high numbers of mid-proportioned variants. The remaining two plots show that having a small proportion of non-random copiers in a population also produces distributions similar to power laws in the lower range of the x-axis. It is possible that empirical data sets of this type may be interpreted as power laws, especially since many techniques in analyzing power laws discard the "noisy" tails (cf. Clauset et al., 2009). However, the tail still remains diagnostic here. Even a small number of conformity copiers are still capable of producing the characteristic long tail composed of a couple remarkably common variants. Although the characteristic hump of mid-frequency variants is not so apparent when only a few anti-conformity copiers are present. These simulations suggest that conformity copiers may be easier to identify in small numbers than anti-conformity copiers.

I will turn now from the social learning perspective on biased cultural transmission purely in terms of behavior copying, and further develop the perspective of cognition-based forces of cultural evolution.

1.7 Inductive evolution

As mentioned earlier, cognition is the locus of cultural change, where cultural variants replicate through a process of reverse engineering. This process is best captured by the concept of inductive inference. In 1739, David Hume first described inductive inference in *A Treatise of Human Nature*, in which he surveyed the essence of being human with the aim of developing an empirical science of human nature. The first book of this treatise, "Of the Understanding", is devoted entirely to human cognition and the types of reasoning that humans naturally engage in. In section III, "Of Knowledge and Probability" he lays out the main form of inference that humans engage in: probabilistic reasoning about cause and effect. He calls this type of reasoning *causal inference* and only in the twentieth century was his concept connected to the modern terminology *inductive inference* by Keynes (1921). Hume describes causal inference as a form of reasoning

in which the relationship among observations (seen or remembered) is imagined or hypothesized:

> ...in all probable reasonings there [must] be something present to the mind, either seen or remember'd; and that from this we infer something connected with it, which is not seen nor remember'd. (Section I.III.VI)

And this is done on the basis of people's past experience:

> Probability ... must in some respects be founded on the impressions of our memory and senses, and in some respects on our ideas. (Section I.III.VI)

Critically, the solutions to inductive problems can be wrong, they cannot be verified by standard logic, and are only supported by various degrees of corroboration from other evidence (Vickers, 2014). Take the following example of inductive inference from Vickers (2014):

> All observed Fs have also been Gs,
> a is an F,

This does not imply "a is a G". Instead, it implies:

> It is probable that, a, not yet observed, is also a G.

Induction can only yield a conclusion with a certain probability and is not an inference process that guarantees a correct inference (Vickers, 2014). In the example above, a is not guaranteed to be G. Because of this, inductive inference lacks proper status as a logic and this shortcoming is discussed by many philosophers as "the problem of induction" (e.g. Popper, 1959; Goodman, 1954; Williams, 1963). However, it is precisely this lack of guarantee in finding the veridical solution that makes induction interesting in a cultural evolutionary perspective. Let me return to the reverse engineering example from the start of this chapter. For a stone tool to be copied, it needs someone to observe its properties, infer how those properties came to be, and then implement them in another piece of stone. A variety of methods can lead to the production of nearly identical stone tools. To copy a particular model one may wonder, were the pieces chipped off with another piece of stone, or with an antler? Was this tool supported against the ground or one's knee while fashioning it? In what order were the different surfaces of the stone chipped? And what is the motor program that should be executed to even chip a stone? If all of the veridical processes that lead to the

creation of the model are correctly inferred, then a very similar copy with result. However, if some are incorrectly inferred, some variations on the model may arise, and these variants will serve as the models for future copying events. Therefore, it is precisely this potential for incorrect inferences that creates the possibility of evolution by induction.

I would like to be able to refer to this source of evolutionary change, so rampant in cultural evolution, not by the concepts of selection or directed mutation, but by the particular biases that cognition imposes on cultural data sets during inductive inference. This brings me to the following definition of inductive evolution:

> **Inductive evolution** is the change over time of entities that replicate via a cognitive process of reverse engineering.

How does inductive evolution relate to other cultural copying mechanisms, such as teaching, emulation, imitation, and overimitation? Reverse-engineering is also central to each of these mechanisms. Let us first look at teaching, which is an event that guides a learner's inductive process. Taking the example of learning to play a viola, a good teacher is able to 1) provide a model of viola playing that maximizes correct inferences in the learner and 2) explicitly correct learner behaviors that will feed into incorrect inferences down the road. However, the bulk of the copying act falls on the learner, where the teacher provides an additional, high-quality source of data to use during inductive inference. Emulation describes the copying of goals actions, without necessarily copying the same actions used to obtain that goal. By definition, reverse-engineering is the successful copying of outcomes, or goals, no matter the route taken to produce the copy. Therefore, emulation describes the subset of reverse-engineering episodes in which new routes (i.e. "incorrect" inferences) are discovered to achieve a certain outcome. Imitation, on the other hand, occurs when both the goal and the actions which lead to the goal are copied, and therefore refers to reverse-engineering via "correct" inferences. And overimitation occurs when actions that have no causal bearing on the goal are copied as well. This might enter into a reverse-engineering process when the goal is not completely understood and the learner is attempting to reverse engineer sub-goals that they can understand.

As for the evolution aspect of inductive evolution, we saw previously that the coarse-grained anatomy of selection and drift processes consist of sampling bias and sampling error. This is also the case for induction and this same coarse-graining applies to probabilistic models of cognition, such as Bayesian models of inductive inference. Bayesian models are stochastic models of cognition that

make inductive inferences by combining prior knowledge about the likelihood of different causal processes in the world, with the likelihood that each of these causes generates certain outcomes, to infer the probability that an observed outcome was generated by each possible cause (Bayes, 1763), but see Jaynes (2003) and Gelman et al. (2013) for the modern formulation of Bayes' rule.

Hume states that the probabilities used during inference rest in our minds, in the form of experience we had with the world and the associations we made between these experiences. These are the probabilities that guide our inductions: they are the knowledge that we bring to a particular problem and constitute everything that goes into making an inductive inference, when faced with a particular data set upon which an inference needs to be made. In Bayesian terms, this prior knowledge is coded as a prior probability over all possible hypotheses (i.e. causes), and can be the result of innate and/or acquired cognitive biases. In Chapter 5, I will describe a particular Bayesian model of frequency learning in thorough detail, with three sub-types, and discuss its two coarse-grained components: sampling bias and sampling error. Furthermore, one of these sub-types has been shown to be equivalent to Wright-Fisher drift with mutation (Reali and Griffiths, 2010). This equivalence clearly highlights the isomorphism between population genetics models and probabilistic models of cognition, in terms of sampling bias and sampling error, and shows that models of inductive inference can support cultural change in a truly evolutionary sense. Even more interestingly, as will be discussed later, models of inductive inference can also support dynamics in cultural evolution that are not equivalent to Wright-Fisher models, and may even define types of evolution that are outside the class of those achievable by molecular evolution.

But for now, I will leave you with a schematic version of induction, which highlights the potential in overlap with models of selective and neutral evolution (Figure 1.15). In this model, an individual may observe a particular distribution over cultural variants (panel a) and produce these variants on ten occasions (panel b). Here, cognition is the source of sampling error and sampling bias and the careful observation of data (panel c) affords estimates of the error and bias that occurs during a particular type of transmission or copying event. In terms of genetic evolution, both selection and mutation (random or directed) are forces of evolutionary change that alter the sampling bias. When a distribution over alleles is observed that seems to be the result of a bias, then the source of that bias is an open empirical question. Likewise, when a distribution over cultural variants appears to be biased in some way, the specific cognitive processes that lead to this bias are also an open empirical question. The entire *inductive loop*, that creates

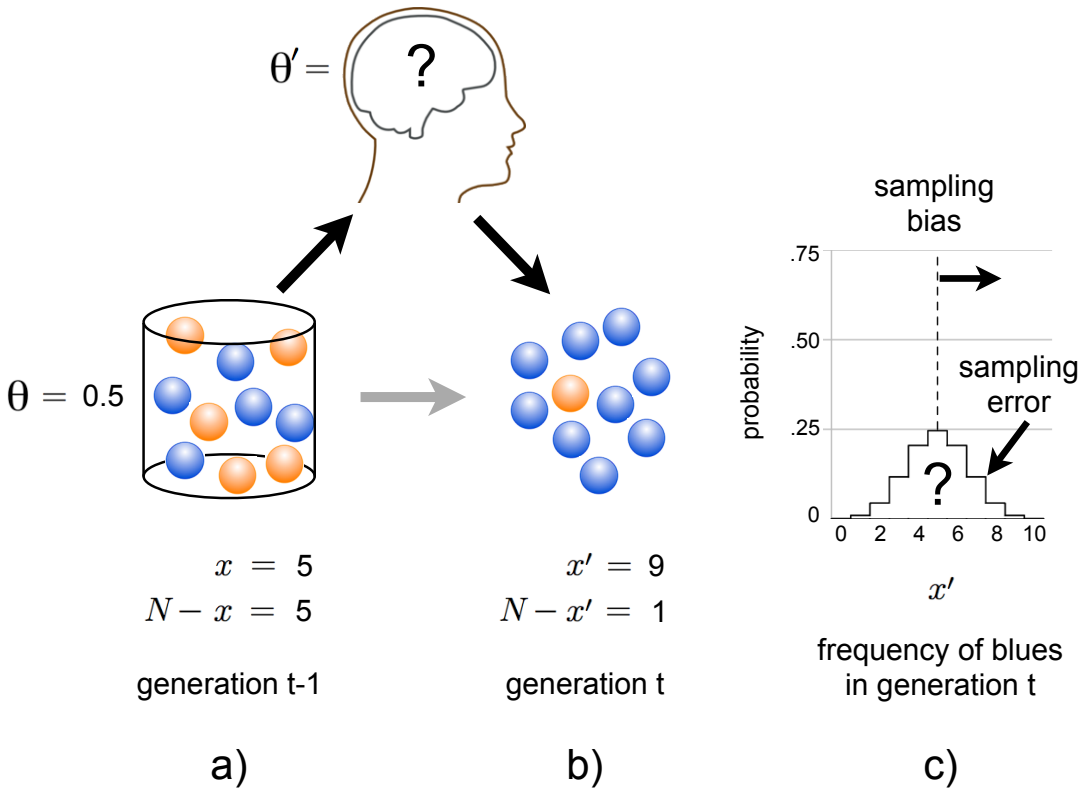


Figure 1.15: This thesis experimentally provides the answer to this schema.

productions from observations, and all of the cognitive mechanisms that allow for the perception, processing, and production of cultural variants, are candidate causes of inductive biases. Although cultural evolution may have isomorphisms to genetic selection, or directed mutation, cultural evolution is driven by inductive biases and these can be understood on their own ground as the driving forces of cultural evolution.

1.8 Borrowing tools for cultural evolution without buying the farm

A general theory of evolution should encompass all forms of evolution that exist. However, our current rich knowledge base of evolutionary theory is very much embedded in its implementations in molecular evolution, population genetics, and the macro-evolutionary processes specific to biological species and ecosystems. Along with these specific implementations, come assumptions which may be inappropriate for cultural evolution. At Darwin's inception of the theory of evolution by natural selection, similar ideas pertaining to cultural evolution were

already in the air (Müller, 1870). But despite similar time depths on scientific recognition of both biological and cultural evolution, a general theory of cultural evolution has not gelled in the social sciences. Mesoudi et al. (2006b, p.330) argue that this is because social scientists are more reluctant than biologists to make simplifying assumptions that abstract away from the complexities of human culture. This reluctance to develop tractable, general models, impedes the accumulation of transferable knowledge within the field, preventing a unifying synthesis among subfields. Mesoudi explains that the evolutionary framework established in the field of biology "brings with it a set of proven methods that have rich potential within the study of culture." Researchers in cultural evolution should make use of established evolutionary theory and borrow from its toolkit to gain insight wherever possible, but at the same time be prepared to push the envelope on biological models, break them to learn what assumptions do not apply to culture, and improve upon these models for their own purposes.

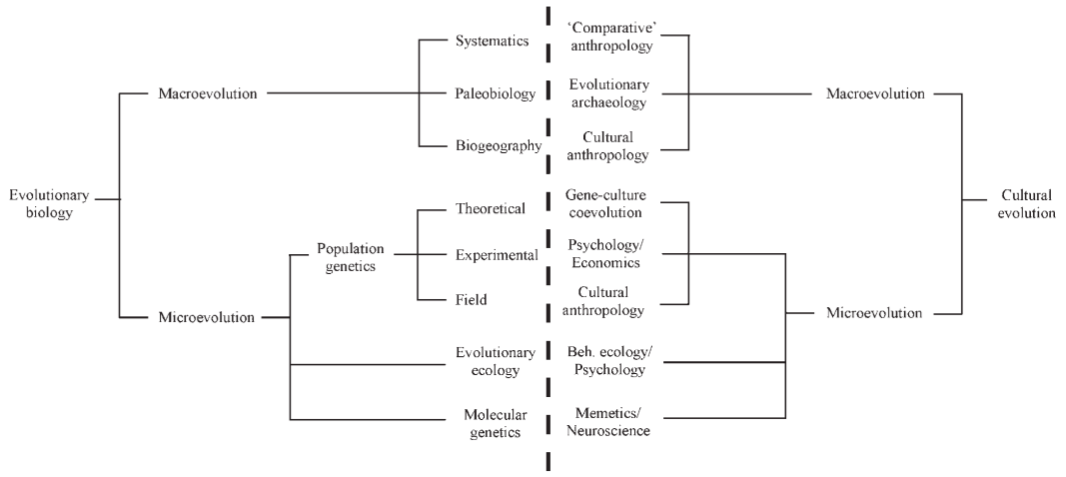


Figure 1.16: The correspondence of the major subdivisions in the fields of evolutionary biology (left) and cultural evolution (right), reproduced from Mesoudi et al. (2006b). The left-hand side of this schema is originally from Futuyma (1998). The interface between experimental psychology and tools from experimental population genetics is where the majority of work in this thesis lies.

Mesoudi et al. (2006b) maps out the subfields regarding cultural and biological evolution, showing correspondences to highlight promising research strategies and potentially fruitful methodological transfer points. For example, Mesoudi places the work of Hahn and Bentley (2003); Bentley and Shennan (2003); Herzog et al. (2004); Bentley et al. (2004), and Bentley (2008), discussed in Section 1.5.2, at the macro-level interface of paleobiology and evolutionary archaeology. Common research goals at this interface include identifying (prehistoric) artifacts, reconstructing their lineages, and determining the evolutionary relationships between

these lineages. Although many of the cultural data sets at this interface are prehistoric and archeological, such as the ≈ 10,000-year-old Paleo-Indian stone projectile points studied by O'Brien et al. (2001), or the 400-year-old West German pottery shards studied by Bentley and Shennan (2003), the fast timescales of cultural evolution also bring recent historical data under this framework. This includes the twentieth-century patent citations, lexical items in academic publications since 1994, and 1990 US census data on first names studied by Bentley et al. (2004) and Bentley (2008), which focus on determining whether these data sets conform to drift or selection at the macro level.

Many examples of cultural macro-evolutionary change that conform to drift are provided by Bentley's research program and he and colleagues conclude that cultural evolution is largely neutral. Although macro-level neutrality does not necessitate that the micro-level copying mechanism is itself neutral (random) copying, it does provide strong evidence for this hypothesis. This takes us to another important interface identified by Mesoudi et al. (2006b): the interface between experimental population genetics and experimental psychology. Research at this level is concerned with the specifics of the copying mechanism, intergenerational change on a short scale, and the dynamics of the transmission process: whether or not certain cultural variants are favored, and why.

Goldstone and Gureckis (2009) argue that "cognitive scientists can leverage our understanding of the limited learning, memory, and decision-making capacities of individuals in order to understand aggregate outcomes" (i.e. the macro level), and they apply a cognitive, micro-level approach to Bentley's baby naming data in a follow-up paper, Gureckis and Goldstone (2009). Here, they model psychological assumptions about frequency information encoding and novelty preferences to provide a closer view of the dynamics on name choices in Bentley's data set. They find that the copying mechanism itself is not likely to be random copying, but a momentum based copying strategy where people seem to track the rate of change in names and choose those which are growing in popularity, relative to others.

A micro-level approach is necessary to state, with certainty, what evolutionary forces underly a particular change that is observed in a data set. And in understanding the forces of cultural evolution, psychological experimentation is a powerful tool because it addresses culture at its locus of change: cognition. Mesoudi et al. (2006b) states:

> Although laboratory-based experiments are an established approach to the study of biological evolution, relatively little experimental work exists in psychology or economics that has studied the dynamics of cul-

> tural transmission. Such studies are essential for a full understanding of cultural evolution. Psychological studies of cultural transmission would benefit from explicitly drawing on the methods of experimental population genetics, both in the design of experiments and in the analysis of data. (p. 339)

This thesis represents a comprehensive attempt to understand the forces of cultural evolution via psychological experiments that were designed to exploit the interface with population genetics models of evolution. In doing so, it adds to the small, but rapidly growing body of experimental cultural transmission research (reviewed in Section 1.3). As recognized in much of this literature, the reverse-engineering process characteristic of induction puts a necessarily cognitive gloss on any attempt to understand the forces, dynamics, and causes of cultural change. Therefore, I will approach this interface with two decidedly cognitive considerations.

First, I will exploit the formal similarity between probabilistic models in population genetics and probabilistic models of cognition to understand the overlap and differences between the forces of biological and cultural evolution. The Wright-Fisher model with and without selection, explained earlier in this chapter, is a probabilistic model with two main components: sampling bias and sampling error. And the Bayesian models mentioned in the previous section (and described in depth in Chapter 5) also define a sampling bias and sampling error on inductive processes.

Second, I will cut close to the operationalization of population genetics models, as the evolution of frequency distributions over variants in a population, and focus directly on cognitive constraints on frequency learning and production. Although Boyd and Richerson (1985) have described the dynamics of frequency-dependent learning, this was done with deterministic models that do not employ sampling error. Given the prominence of neutral change as an explanation for macro-level cultural evolution (e.g. Bentley et al., 2004), we need to understand how neutral change could occur via the inductive task of frequency learning and reproduction.

All of the experimental work presented in this thesis is aimed at describing the sampling bias and sampling error involved in frequency learning for one of humankind's most remarkable culturally-transmitted data sets: language. A great deal of literature exists on the tracking, processing, and production of language on the basis of its frequency information (e.g. Jescheniak and Levelt, 1994; Redington and Chater, 1996; Elman, 1998; Saffran, 2003; Lieberman et al., 2007; Hudson Kam and Newport, 2009). In Experiment 1, I begin by exploring human sampling error during a basic inductive task of frequency learning and ask

if human inductive inference can support cultural drift. Then, in a further 4 experiments, I will extend this basic frequency learning experiment with conditions that engage different cognitive biases in frequency learning. The particular class of biases I will focus on is regularization: how people eliminate variation in language. Then, in Chapter 5, I fit a Bayesian model of frequency learning and production to the data sets from each experiment to provide a window on the inductive biases operating within participants' heads as they produce the biased behavior that they do. In essence, what each of these experiments and models do it provide a solution to Figure 1.15, in its own way. Last, in Chapter 6, I analyze the cultural evolution of these frequency distributions by extrapolating the experimental data forward over several generations of learners. This provides a view on the evolutionary dynamics associated with different inductive biases in frequency learning.

Chapter 2

A cognitive basis for cultural drift

2.1 Experiment 1: probability matching in frequency learning

At the macro level, cultural evolution often seems to conform to drift (Bentley et al., 2004). Because cultural variants are never directly copied, but reverse engineered via a processes of inductive inference, we need to know what cognitive mechanisms can support cultural drift. Such a mechanism would need to reproduce the relative frequencies of all the observed variants with a form of error that is not biased toward any one of those variants. If the relative frequencies were reproduced veridically, without error, then this would preclude the cultural evolution of this data set because the distribution over cultural variants would never change. If the error were biased toward any particular variant, that variant would be more likely to replicate than the others, and this would not be a true drift process (i.e. constitute neutral evolution).

This experiment was designed to explore probability matching behavior as a possible basis for drift in cultural evolution. Probability matching is a well-studied human decision making strategy and basic frequency learning behavior that leads humans to replicate the relative proportions of events that they observe with high fidelity, but some amount of error. Most studies in probability matching are interested in how the majority of participants behave (i.e. the mode and mean of participant responses) and thus, not much is documented about the particular variance on error associated with probability matching.

People are very good at tracking the frequencies of events in their environment and can use this information to make predictions about future events (e.g. Hasher

and Zacks, 1979; Gelman, 1998). Much work in the field of psychoeconomics has described probability matching as a sub-optimal decision-making strategy that people tend to employ when using past frequency information to predict future events (see Vulkan (2000) for a review of 22 such experiments). However, this sub-optimality isn't necessarily because people aren't rational, it is because they do not fully understand the nature of random events (Bar-Hillel and Wagenaar, 1991) and seem to be hypothesizing rules that govern the sequence of events (Wolford et al., 2004; Gaissmaier and Schooler, 2008; Unturbe and Corominas, 2007). The rational strategy for making predictions about truly random events is maximizing: choosing the most frequent item all of the time. Say, for example two events (a, b) occur with the probabilities (0.3, 0.7). In this case, a maximizer will bet on event b all of the time and be correct 70% of the time ($0.3 \cdot 0 + 0.7 \cdot 1 = 0.7$), but a probability matcher would only be correct 58% of the time, because they are expected to choose each event with the probability of its occurrence ($0.3 \cdot 0.3 + 0.7 \cdot 0.7 = 0.58$).

When probability matching behavior is obtained in artificial language learning experiments, it also seems to be the result of a pattern search. Perfors (in preparation) shows that modulating the extent to which participants believe that observed linguistic variation is meaningful modulates their reproduction of that variation. In one condition, adult learners participated in a typical artificial language learning task and probability matched by reproducing the linguistic variants in the same frequencies they observed them. In a second condition, participants learned the same artificial language, but were made to believe that the variation is the result of errors made by other learners of the artificial language. In this condition, their behavior moved more toward maximization: they tended to produce the most frequent variant with a higher frequency than they observed it. (This is known as *regularization* in the language learning literature and is similar to maximization. The formal overlap of regularization and maximization is explained in Chapter 4.)

Research on probability matching is split in terms of how it defines this behavior. Some work on the subject visually inspects the mode or mean of participant responses for behaviors that were nearly identical to the input frequency, or in time course studies, oscillated about the mean. (Edwards, 1961; Kirk and Bitterman, 1965; Herrnstein, 1970; Miller and Valsangkar-Smyth, 2005). Other experiments that define probability matching behavior more rigorously, look for mean participant behavior that is not significantly different from the input frequency (Wolford et al., 2000; Shanks et al., 2002; Hudson Kam and Chang, 2009; Hudson Kam and Newport, 2009; Reali and Griffiths, 2009; Smith and Wonnacott, 2010;

Culbertson et al., 2012). However, occasionally, bimodal responses in which the mode is not on the input frequency, can yield a mean that is not significantly different from the input frequency. This would describe probability matching behavior at the population level (e.g. Reali and Griffiths, 2009), but not at the individual level, which is the level that an investigation of the inductive processes associated with unbiased sampling is concerned with. Therefore, throughout this thesis, I will use the following definition of probability matching behavior:

> **Probability matching** is when the mode of participant responses is on the input frequency and the mean is not significantly different from the input frequency.

Additionally, there are several identifiable sub-types of probability matching behavior that I will refer to occasionally in the following chapters.

1. *Perfect probability matching.* The mode of participants' responses is on the input frequency and there is no error.

2. *Imperfect probability matching.* The mode of participants' responses is on the input frequency and there is some error. This error can be specified, for example, as in *binomial probability matching*, in which the distribution of participants follow a binomial distribution, *gaussian probability matching*, in which the distribution of participants follow a truncated gaussian distribution with a particular variance, or any other distribution that characterizes the distribution of participant responses about the mean. This error can be biased or unbiased, leading to the following two subclasses of imperfect probability matching:

3. *Biased probability matching.* The mode of participants' responses is on the input frequency, but the veridical mean is not. For finite samples, this process can end up under the broad definition of probability matching when the sample shows no significant difference between participants' mean and the input frequency. This constitutes a biased sampling process that may still be referred to as probability matching behavior in the empirical literature.

4. *Unbiased probability matching.* The mode of participants' responses is on the input frequency and the veridical mean is also on the input frequency. This type of probability matching is a form of unbiased sampling error which, again, can be specified by any number of distributions in which the mode equals the mean. This is the type of probability matching behavior

that is (hopefully) confirmed when the mean of participant behavior is found to be not significantly different from the input frequency.

To date, there has not been a study that explicitly investigates the plausibility of human probability matching behavior as a basis for neutral cultural evolution. This experiment consists of a basic frequency learning task in which participants observe several blue and orange marbles being drawn from a bag and are then asked to demonstrate some more draws that are likely to come out of that bag. This task involves induction, in the sense that participants infer the relative proportion of the two colors in the bag on the basis of limited data and then arrive at a new sample of data. A variety of biases may be at play during this inductive cycle, such as participants' biases in the perception and production of random sequences, their a priori expectations about the relative proportion of marbles in bags, and any biases related to representativeness (since participants are being taught about these marbles and in turn, have to demonstrate what they've learned afterward). The goal of this experiment is to assess whether or not the cognitive processes involved in frequency perception, processing, and production can lead to a globally unbiased sampling process that would support cultural drift.

2.1.1 Method

Participants

238 participants were recruited via Amazon's Mechanical Turk crowdsourcing platform and completed our experiment online. Participant location was restricted to the United States of America and verified by a post-hoc check of participant IP address location. 46 participants were excluded on the basis of the following criteria: self-reporting the use of a pen or pencil during the task[1] (11), not reporting their sex or age (2), or having previously participated in this or any of my experiments, as determined by their user ID with MTurk (11). More participants were recruited than necessary with the expectation that some would be excluded by these criteria. Once the predetermined number of participants per condition was met, the last participants were excluded, totaling 22 participants across all conditions. All excluded participants received the full monetary reward for the task, which was 0.10 USD. The average time taken to complete the experiment was 3 minutes and 38 seconds, with a standard deviation of 1 minute

[1] In an exit questionnaire.

and 19 seconds. Of the final 192 participants, 64% are female and the mean age is 36.0 (min = 18, max = 68) with a standard deviation of 11.9 years.

Materials & Stimuli

The experiment was coded up as a Java applet that ran in the participant's web browser in a 600x800-pixel field. Coding was done with Processing[2] and the server was written in Python[3]. Stimuli consisted of a photograph of a pouch and graphically generated images of marbles in 2 colors: blue and orange (Figure 2.1). The RGB color of the blue marble is 0, 51, 204 and the orange is 255, 128, 0. This pair of colors was chosen to be distinctive (i.e. their hues lie on approximately opposite sides of the color wheel) and because they can also be easily distinguished by individuals with various forms of color blindedness (Rigden, 1999).

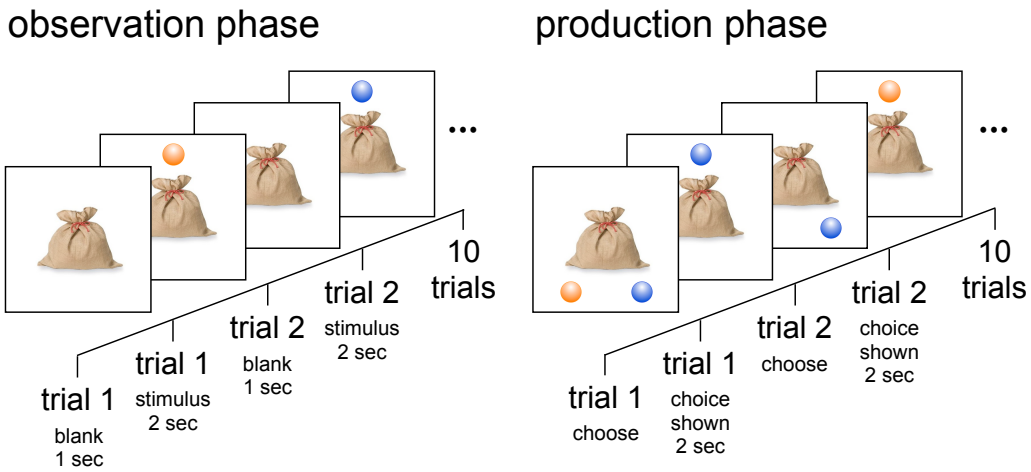


Figure 2.1: Schema of the observation and production phases for Experiment 1.

Procedure

The experiment consisted of an observation phase and a production phase, with 10 observation trials and 10 production trials (Figure 2.1). In each observation trial, the pouch was displayed on its own for 1 second and then a marble was displayed above it for 2 seconds, with no break between trials. In each production trial, the pouch was displayed and the two marbles were displayed below it near the bottom left and right corners of the screen. Each production trial had no time constraint. When participants scrolled over one of the marbles, a grey box appeared around that marble. When the participant clicked on the marble,

[2]www.processing.org
[3]www.python.org

the box turned bold, the marble they selected appeared above the pouch for 2 seconds (in the same location as it had during the observation phase) and that marble color was registered as the participant's response for that production trial. Production trials repeated until 10 responses were collected. Participants were not told how many observation or production trials there would be. Complete instructions and exit questions can be found in Appendix A.

Conditions

All possible ratios for 10 draws were tested. Participants were randomly assigned to one of 6 observation ratios (x:y) = 5:5, 6:4, 7:3, 8:2, 9:1 and 10:0. Here, x corresponds to whatever color was the *majority* marble during the observation phase and y corresponds to whatever color was the *minority* marble during the observation phase. Color was counterbalanced so that half of the participants in each observation ratio saw blue as the majority marble and half saw orange as the majority marble. From here on, whatever marble color corresponds to x will be referred to as *observed majority* and whatever marble color corresponds to y will be referred to as *observed minority*. The test-side presentation of marbles was also counterbalanced so that half of the participants had their *observed majority* as the right-hand test choice and half had their *observed majority* as the left-hand test choice. This yields 4 counterbalance conditions:

1. *observed majority* is blue and on test-right
2. *observed majority* is blue and on test-left
3. *observed majority* is orange and on test-right
4. *observed majority* is orange and on test-left

Overall, two counterbalance manipulations and six observation ratios yields a 6x2x2 experimental design, with 8 participants in each group.

2.1.2 Results

Figure 2.2 (top row) shows the results of Experiment 1. Each column corresponds to one of the six observation ratios, ranging from 5:5 (on the left) to 10:0 (on the right). Each pane contains the distribution of ratios that participants produced in response to one observation ratio. These production ratios are displayed on the x-axis as the number of times a participant produced *variant x* from the observation ratio *x:y*. *Variant x* corresponds to whatever marble/word was in

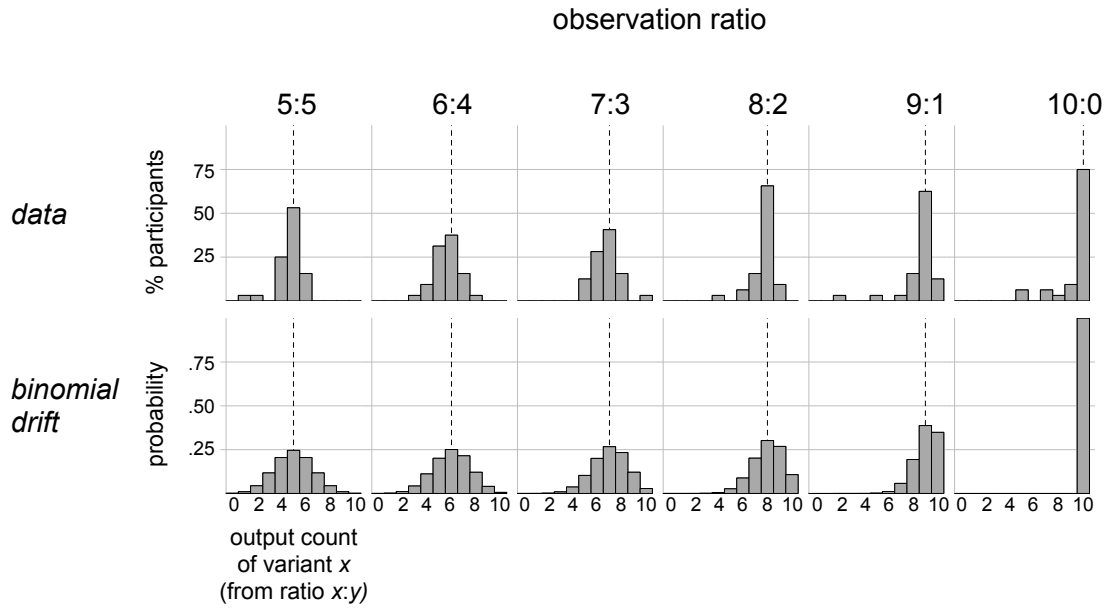


Figure 2.2: Top: The results of Experiment 1. Bottom: The distribution of behaviors that would be obtained in one generation of binomial drift.

the majority during the observation phase.[4] All observation ratios are indicated by a dashed line. For example, the leftmost panel gives the results for the 32 participants who observed a 5:5 ratio of marbles. Here, we see that 53% of these participants also produced a 5:5 ratio, 25% produced a 4:6 ratio, 16% produced a 6:4 ratio, and no participants produced only one marble color (i.e. a 10:0 or 0:10 ratio).

Participants appear to be probability matching with some error. The mode of participant behavior is on the input frequency, however some participants over-produced the majority marble (responses to the right of the dashed line) and others under-produced the majority marble (responses to the left of the dashed line). One-sample t-tests were conducted to determine if the participants' mean production of the *observed majority* differed significantly from its observed frequency. A two-tailed test was used because there was no a priori hypothesis about the directionality of this potential difference. There was no significant difference in observation ratio conditions 5:5, 6:4, 7:3, 8:4, and 9:1, however some of these p-values are quite low: (5:5, $t(31) = -1.6667, p = 0.11$; 6:4, $t = (31) - 1.9823, p = 0.06$; 7:3, $t(31) = -1.4669, p = 0.15$; 8:2, $t(31) = -1.8317, p = 0.08$; 9:1, $t(31) = -1.6253, p = 0.11$). In the 10:0 condition, mean production of the *observed majority* is significantly lower than its observed frequency ($t(31) =$

[4]In the 5:5 observation ratio there is no majority variant, so a random marble/color was coded as *variant x*.

Counterbalance condition:	1	2	3	4
observed majority responses	335	350	342	337

Table 2.1: The number of production trials where participants chose the *observed majority* marble is similar across all counterbalance conditions.

$-2.6417, p = 0.01$). This significant result may be due to the extreme ceiling effect of the x = 10 condition, because noise on participant responses can only lower the mean. Participant behavior in this experiment falls under the definition of probability matching behavior because the model participant responses fall on the input frequency, and the mean of their responses are not significantly different from the input frequency.

Given the robust probability matching profile, it seems that input proportion is the main predictor of participant behavior in this experiment. However, there could be an effect of the two counterbalance manipulations: color of the majority marble and test side. And there could be an interaction between these conditions, for example the *observed majority* marble might be more likely to be chosen when it is blue *and* the right-hand test choice. Table 2.1 gives the raw number of trials in which participants responded with the *observed majority*, per counterbalance condition. Here, we see a near even divide: participants' likelihood of choosing the *observed majority* is similar across all counterbalance conditions. This indicates that there is no effect of counterbalance condition of on the frequency with which participants produce the *observed majority*.

To assess statistical significances in the relationship between the frequency with which participants produced the *observed majority* marble and the observation ratios and counterbalance conditions, I performed a linear regression analysis in R (R Core Team, 2013) with the *lme4* (Bates et al., 2013) package. The dependent variable was the production frequency of the *observed majority* marble. The independent variables were 1) observation frequency of the *observed majority*, 2) color of the *observed majority*, and 3) right vs left test side location of the *observed majority*. All independent variables were entered with interaction terms. There was a significant effect of observation frequency on the production frequency of the majority marble ($t(184) = 8.636, p < 0.001$), as would be expected because participants are probability matching and have clearly learned about the observation ratios. Neither of the counterbalance conditions significantly predicted participants' responses: color was not significant ($t(184) = -1.107, p = 0.27$) and test side was not significant ($t(184) = 0.211, p = 0.833$). Likewise, there were no significant interactions between any of the three independent variables[5]. This

[5]observation ratio and color: $t(184) = 1.271, p = 0.21$, observation ratio and test side:

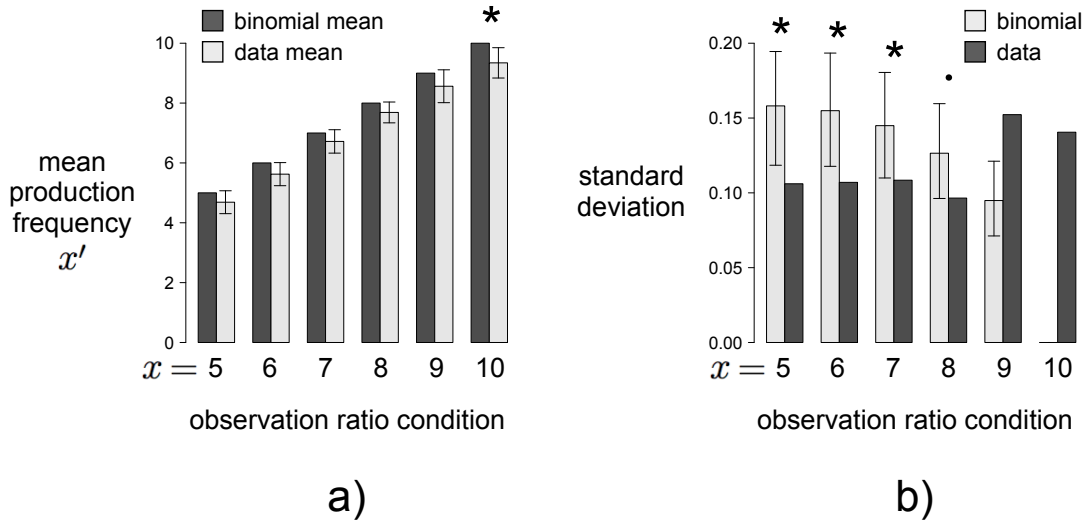


Figure 2.3: Comparing the probability matching data and binomial drift for $N = 10$, per observation ratio $(x : y)$ in terms of a) mean and b) variance.

means that participants showed no bias toward one of the marble colors or one of the test sides: observation frequency alone predicts participants' performance.

Comparison to binomial drift

How does this probability matching profile compare to binomial drift? Figure 2.2 (bottom row) shows the distribution of behaviors that would be obtained in one generation of binomial drift (refer back to Section 1.5.2). These binomials are the distributions we would expect participants to produce if they were doing something isomorphic to randomly sampling their production trials, with replacement, from their observation trials. Through visual inspection, the results of Experiment 1 look similar to the binomial distributions, maintaining the mode on the input proportion. However, the means appear to be slightly biased toward the minority variant and the variance appears restricted. First, let's discuss the means. In these binomials, the mean falls on the observation frequency, x. Because of this, the t-tests also tell us that there is no significant difference between mean participant behavior and binomial means for the 5:5, 6:4, 7:3, 8:4, and 9:1 observation ratios. This is important because it tells us that this probability matching behavior might support neutral evolution, because this is a necessary condition for neutral change: neither variant should be privileged. Figure 2.3a plots the binomial mean against the empirical mean for each of the six obser-

$t(184) = 0.075, p = 0.94$, color and test side: $t(184) = 0.248, p = 0.81$, observation ratio, color and test side: $t(184) = -0.529, p = 0.60$.

vation ratio conditions. The error bars are 95% confidence intervals indicating where the true mean of participant responses lies. Here we see that all of the data means seem to be on the low side. Although these values are not significantly lower than the binomial means, it is possible that a larger participant pool may reveal a consistent bias toward the *observed minority* variant.

The variance, on the other hand is significantly different between participant behavior and binomial drift, and this confirms that cultural drift via probability matching behavior is not equivalent to binomial drift. Figure 2.3b plots the standard deviation[6] of participant responses against the standard deviation of the matched binomials from Figure 2.2b. To determine whether the standard deviation of participant responses is significantly different from that of its matched binomial, we need to find out how likely it is that binomial sampling could generate a standard deviation that is at least as extreme as the one generated by participants. For each observation ratio condition, I ran a Monte Carlo analysis that sampled 32 data points from the matched binomial, computed the standard deviation of these data points, and repeated this 1024 times[7]. The resulting 1024 standard deviations were ranked and these 95% confidence interval and p-value of the empirical standard was determined on the basis of this ranking.[8] The error bars in Figure 2.3b show the 95% confidence intervals. The standard deviations in conditions 5:5, 6:4, and 7:3 were all significantly lower than that of the binomial (5:5, $p = 0.006$, 6:4, $p = 0.012$, 7: 3, $p = 0.04$). For 8:2, the standard deviation is close to significance ($p = 0.051$). Standard deviations were significantly higher in conditions 9:1 and 10:0 (9:1, $p = 0$; 10:0, $p = 0$)[9].

How does this difference in variance affect the dynamics of drift? Neutral copying with low variance maintains variation in the population longer. This was discussed at the end of Section 1.5.1 in relation to population size: larger population sizes produce binomial drift distributions with lower variance and thus, variation loss (i.e. fixation) takes a longer time on average to occur. The numerical calculation of average fixation time for binomial drift (Figure 1.3) was repeated for probability matching using the raw data in Figure 2.2 as the transi-

[6] which is the square root of the variance

[7] Singh and Xie (2008) recommend a total number of re-runs equalling the square of the number of participants ($32^2 = 1024$).

[8] Here, the 25th lowest value is the lower bound on the 95% confidence interval and the 947th lowest value is the upper bound. This is a two-tailed test. The p-value is determined by finding the rank of empirical value, dividing it by the total number of ranked items, and multiplying by two to get a two-tailed p-value. For example, if the empirical value corresponds to the 30th lowest item, its p-value will be $\frac{30}{1024} = 0.059$.

[9] These p-values are zero because there were no instances of the model producing a standard deviation more extreme than the empirical value.

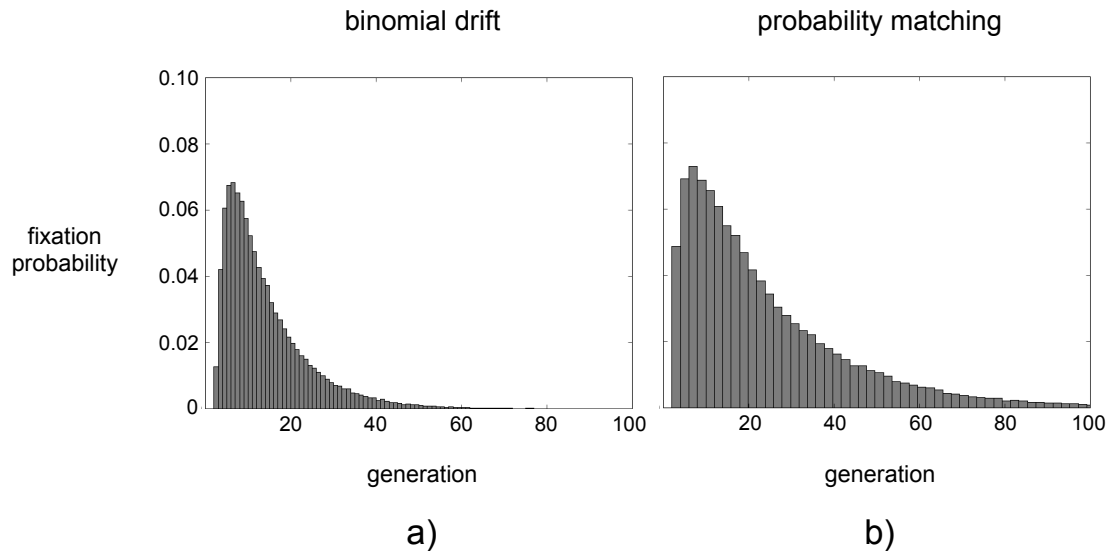


Figure 2.4: Fixation time probability distributions. a) Binomial drift where $N = 10$, mean fixation time is 13.86 generations. b) The raw probability matching data obtained in Experiment 2 (also $N = 10$), mean fixation time is 24.17 generations. These were numerically calculated from 100,000 independent evolutionary trajectories initialized at $p = 0.5$.

tion probabilities $x \rightarrow x'$. Figure 2.4a re-plots the fixation probability distribution for binomial drift where $N = 10$ and Figure 2.4b plots the distribution for the probability matching data, where $N = 10$ because this equals the population size of cultural variants (blue and orange marbles) that participants can observe and produce in the experiment. The average generation at fixation for the probability matching data is 24.17 generations. This is much longer than $N = 10$ binomial drift, which has an analytical solution of 13.86 generations. In fact, the binomial model that comes closest to the probability matching data result is a larger population size, $N = 17$ with average fixation at 23.57 generations.

2.1.3 Discussion

The results of this experiment show that human probability matching capabilities are a plausible basis for cultural drift. Three types of sampling bias could be at play in this task, however no evidence was found to support them: 1) Participants did not possess a direct bias for a variant on the basis of its color. 2) Participants did not possess a production bias on the basis of test side location. 3) Participants did not demonstrate a significant frequency-dependent bias that would lead them to over-produce the majority or minority variant. This means that the probability matching behavior obtained in this task constitutes an unbiased, neutral sampling

process, which would necesarily lead to the neutral evolution of marble frequencies over time. However, probability matching behavior in this task was not identical to binomial drift because the variance in output frequencies was significantly lower. The fact that participants' output frequencies displayed a significantly lower variance than that of binomial drift means that the cultural drift supported by probability matching is not binomial: it is not equivalent to the Wright-Fisher model of neutral evolution. One likely reason for this restricted variance is that participants are trying to be representative. That is, they are producing responses that have high likelihood under their estimate of the ratio of marbles in the bag. For example, if you inferred a 7:3 ratio of marbles in the bag, and were asked to tell someone what a likely set of draws from this bag looks like, you may try to produce the most representative set, which is also a 7:3 ratio. This type of data would give them the best shot at inferring the the same ratio that you had inferred. Representativeness of this type has been explored experimentally (e.g. Kahneman and Tversky, 1972; Grether, 1992) and in cognitive modelling (e.g. Tenenbaum and Griffiths, 2001; Rafferty and Griffiths, 2010).

One consequence of this restricted-variance form of cultural drift is that cultural drift may eliminate variation at a slower rate than genetic drift does. This has important implications for the use of drift models as null hypotheses, especially in relation to regularization studies because they deal with the selective elimination of variation over generations of learners. A good indicator that a trait is undergoing selection is that it fixes in a population significantly quicker than it would under drift. If cultural drift eliminates variation at a much slower rate than genetic drift does, that means that the use of binomial drift models (imported directly from models of genetic evolution) will make the detection of a wide range of weak selection pressures impossible for cultural data sets. For example, let's suppose a certain cultural selection pressure operates in accordance with the basic selection model described in Section 1.6.1, Equation 1.3. If we assume that the sampling error on this model is not binomial, but instead is defined by a Gaussian distribution with the same standard deviations[10] as the data obtained in Experiment 1, the entire range of selection strengths from about $s = 0.27$ down to $s = 0$ would not eliminate variation faster than binomial drift would and therefore, would not be correctly detected as selection if a binomial

[10]The standard deviation values obtained per observation ratio were used to define the sampling error about the model's expected values, θ'. (Refer back to the plots in Figure 1.8, which define the selection function that maps θ to θ'.) For $\theta = 0.5$, the standard deviation of the data in the 5:5 condition was used, for $\theta = 0.6$, the standard deviation from the 6:4 condition was used, and so on. The exact standard deviations, graphically shown in Figure 2.3 are as follows: 5:5 = 0.1061, 6:4 = 0.1070, 7:3 = 0.1085, 8:2 = 0.0965, 9:1 = 0.1523, 10:0 = 0.1405.

drift baseline were used as a null hypothesis. In this model, $s = 0.27$ has an average fixation time of about 13.8 generations, close to binomial drift at 13.86, whereas $s = 0.28$ is a little lower at 13.5 generations, and $s = 0.26$ is a little higher at 14.2 generations. If restricted-variance drift is a general characteristic of cultural evolution, then the many data sets which Bentley and colleagues have shown to conform to binomial drift may very well be the result of non-neutral copying processes. This suggests an important avenue for future research: we need to understand what effects variance-restricted neutral sampling processes have on macro-level analyses of cultural evolution, such as power law distributions. Do these processes also produce power law distributions, or do they result in detectable deviations from the power law?

Finally, what is the reason that sampling error in cultural drift is so low? In this basic frequency learning task, participants observe orange and blue marbles in a particular ratio, and then they produce them in a particular ratio. These variants are not being copied *directly*, they are passing thorough participants' minds and their replication hinges on the cognitive mechanisms involved in perception, processing, and production. The 10 variants that each participant observes lose their discreteness when they pass through cognition. People may be forming representations of these variants and their relative frequency. In the case of probability matching, participants are able to reproduce the observed ratio *better* than a random sampling process such as drift, or an exemplar model that stores all variants with their veridical frequencies and generates productions by randomly sampling these variants, with replacement, from perfect memory. Instead, participants in this task are employing a form of inductive inference to reverse-engineer 10 draws that are likely to come out of a bag, on the basis of 10 draws they have observed from that bag. The results of this experiment show that the inductive inference process can be less error prone than a random sampling process, and in turn, lead to higher levels of transmission fidelity than processes that sample directly from populations of cultural variants.

In short, there is no reason to assume that the mathematics of sampling physical entities in the world, or enumerable observations of cultural variants, should apply to the type of sampling error that occurs in cultural evolution processes. A better understanding of the cognitive bases of cultural drift is needed to provide cultural evolution research with appropriate null hypotheses for the accurate detection of bias-driven causes of cultural change.

Chapter 3

The psychology of regularization in language

In the previous chapter, probability matching was identified as an unbiased frequency learning behavior that could support neutral drift in cultural evolution. From this point forward, I will explore the cognitive biases that take participants away from probability matching and cause them to eliminate variation during frequency learning. This selective elimination of variation is known as regularization, and has been studied in depth in the field of linguistics. Human languages contain very little unpredictable variation (Chambers and Schilling-Estes, 2013) and when language learners do encounter variation, they tend to regularize it. Regularization has been documented extensively in natural language use contexts, such as the formation of Creole languages from highly variable Pidgin languages (Bickerton, 1981; Sankoff, 1979; DeGraff, 1999; Lumsden, 1999; Meyerhoff, 2000; Becker and Veenstra, 2003), the formation of new signed languages (Senghas et al., 1997; Senghas, 2000; Senghas and Coppola, 2001), historical trends of language change (Schilling-Estes and Wolfram, 1994; Lieberman et al., 2007; van Trijp, 2013), and children's acquisition of language (Berko, 1958; Marcus et al., 1992; Singleton and Newport, 2004; Smith et al., 2007) Regularization has also been documented extensively in the laboratory through artificial language learning experiments (Wonnacott and Newport, 2005; Hudson Kam and Newport, 2005, 2009; Reali and Griffiths, 2009; Smith and Wonnacott, 2010; Perfors, 2012; Culbertson et al., 2012). Because I will be taking an experimental approach, with adult learners, to understanding non-neutral cultural evolution via regularization biases, I will take a moment here to summarize some key existing regularization experiments with adult learners.

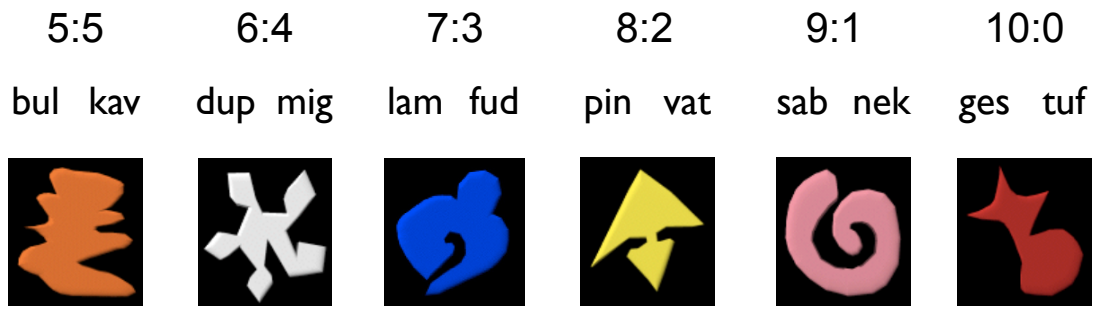


Figure 3.1: The stimuli used in Reali and Griffiths (2009). Top: the relative proportions of each word pair. Middle: the word pairs. Bottom: the objects (these were animated to appear 3-dimensional and shown as videos).

Reali and Griffiths (2009) address regularization at the lexical level and investigate how participants eliminate variation among word-meaning mappings. Participants were trained on an artificial language that contained six meanings (6 different objects), where each meaning was paired with two possible words (two synonyms for that meaning). Figure 3.1 shows the object and word stimuli they used.[1] Participants observed 10 naming events per object, totaling 60 training trials. On each trial a random object was chosen and it was named with one of the two possible words. Stimuli were presented visually and auditorily. Each of the two words in a pair was shown in a particular relative proportion to one another. Each of the six word pairs followed one of the six possible relative proportions for 10 trials: {5:5, 6:4, 7:3, 8:2, 9:1, 10:0}. Because these mappings are not all deterministic, one-to-one mappings between words and meanings (i.e. they aren't all 10:0 ratios), this stimuli set contains unconditioned variation among lexical items. And each ratio differs in its variability, with 5:5 ratios being the most variable, 6:4 a little less variable, 9:1 only slightly variable, and 10:0 not variable (i.e. fully regular).[2] Then, in a testing phase, participants were shown an object along with the two words that had appeared with it during training. Participants were instructed to select one of the words. Each object was shown 10 times in the testing phase, totaling 60 testing trials. Therefore, 10 naming events were collected per object. The data of interest here were the relative proportions of the words pairs that participants produced, and how they differed from the relative proportion in which they were observed.

Figure 3.2 shows the results of Reali and Griffiths (2009). In the panel to the right we can see that many participants regularized their productions by only

[1] These are the exact word stimuli, but the pairings among words may not be the same to those used in the experiment. These images were kindly provided by Florencia Reali.

[2] The upcoming Section 3.1 explains how these different levels of variability can be quantified.

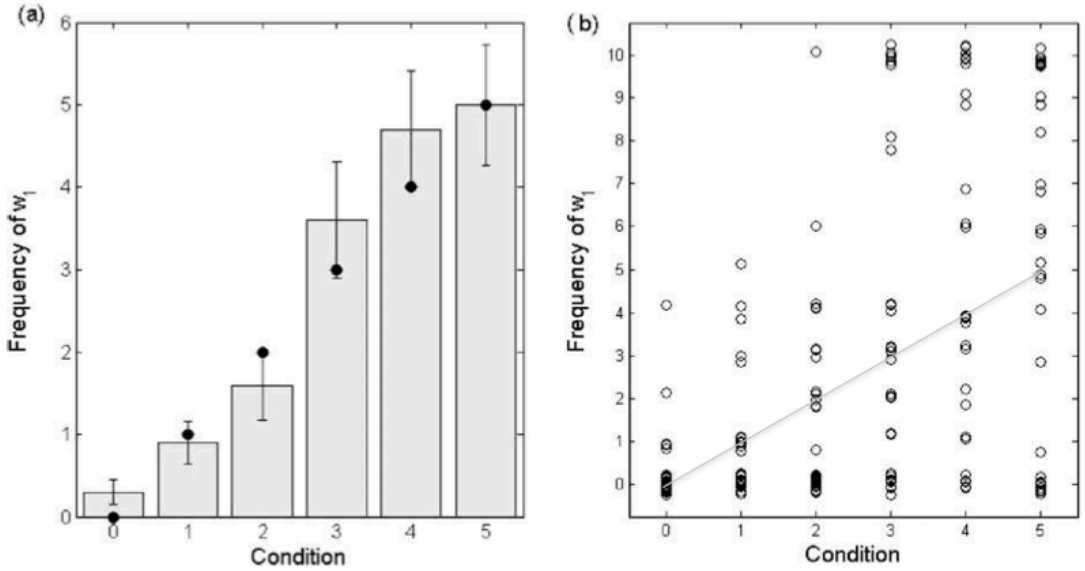


Figure 3.2: The lexical regularization results of Reali and Griffiths (2009). Each condition corresponds to one of the training ratios. Condition 1) 10:0, 2) 9:1, 3) 8:2, 4) 6:4, 5) 5:5. Left: mean frequency of word y from the training ratio, *x:y*. Black dots correspond to the frequency of word y in the training stimuli and error bars indicate one standard error of the mean. Right: The individual data points per participant.

producing one of the words per object (with a frequency of $y = 0$ or $y = 10$). More participants appeared to regularize by overproducing the word that occurred most frequently in the training stimuli (word x from the training ratio *x:y*). However, some participants seem to probability match, because there are production frequencies of y that equal their observed frequency during training (these are the data points that fall on the gray line, which I have added to the figure). The panel to the left shows the mean of participants' production frequencies per condition. According to the most common definition of probability matching in the literature, this result constitutes probability matching behavior because there is no significant difference between participants' mean production frequency of word y and the observed frequency of word y. However, this definition is misleading because the mode of participant behavior in this data set is not on the mean: it is bimodal and participants are clearly regularizing their responses. Reali and Griffiths (2009) do adopt the standard definition here and therefore incorrectly characterize these responses as probability matching behavior. It is precisely this potential for misidentifcation which prompted the more stringent definition of probability matching behavior which I laid out in Section 2.1.

Reali and Griffiths (2009) also conducted this experiment in an iterated learn-

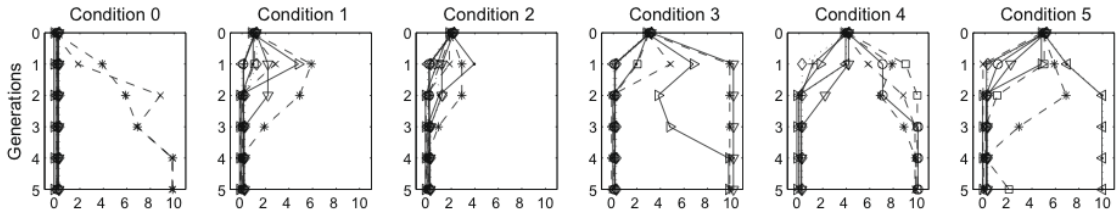


Figure 3.3: The results of the iterated learning experiment in Reali and Griffiths (2009). Each line shows the trajectory of one iterated learning chain for five generations. The x-axis shows the produced frequency of word y from the initial training ratio *x:y*.

ing format (refer back to Section 1.3), in which the production ratios of one participant served as the training ratios for the next. When word frequencies are culturally transmitted, they quickly converge toward fully-regular ratios (10:0 or 0:10). Figure 3.3 shows the results of the iterated learning chains. Each chain was initialized in one of the six possible training ratios. These results show that individual participants possess some sort of regularization bias for one-to-one mappings between words, but when their results are culturally transmitted, this bias becomes very clear, as participant behavior converged to 10:0 production ratios in only a few generations of learners. This difference between single-generation behavior (obtained from the typical experimental design that compares one set of training and testing data) and the behavior after multiple generations of learners (obtained from the iterated learning experimental designs) exemplifies the use of iterated learning as a tool for revealing inductive biases and will be further discussed in Chapter 6.

Hudson Kam and Newport (2009) investigated regularization of determiners. In the first experiment in this paper, participants were trained on an artificial language consisting of 36 nouns, 7 intransitive verbs, 5 transitive verbs, 1 negative marker, and 2 main determiners, one for each of the two noun classes. Each training trial consisted of a video of a scene and a sentence in the language that described that scene. Sentences were presented auditorily only. Testing trials consisted of a spoken sentence completion task in which participants saw a novel video of a scene and heard the first word (always a verb) of the corresponding sentence. There was one control condition and four experimental conditions. In every condition, the main determiner occurred in 60% of the training sentences. In the control condition, no determiner occurred in the remaining 40% of the sentences. In the experimental condition 2-ND, two additional *noise* determiners occurred in the remaining sentences (20% with one noise determiner and 20% with the other noise determiner). Condition 4-ND there were 4 noise determiners,

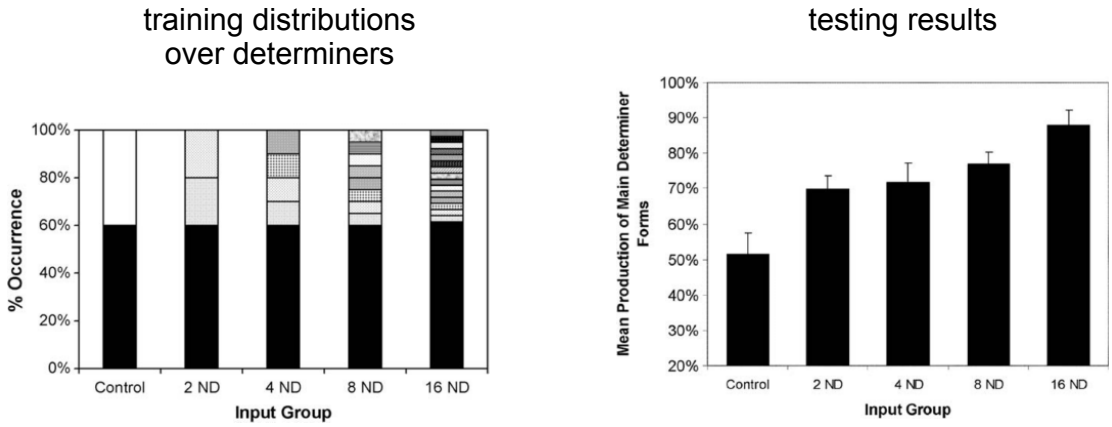


Figure 3.4: Results of Hudson Kam and Newport (2009). Left: training distributions over determiners in each of the five conditions. Black indicates the percentage of occurrences of the main determiners and grays indicate the different noise determiners. In the control condition, there were no noise determiners. Right: mean number of productions of the main determiner in each of the five conditions.

in condition 8-ND there were 8, and in condition 16-ND there were 16. Noise determiners were always presented with an equal number of sentences (Figure 3.4, left). The results of the sentence completion trials are shown in Figure 3.4 (right) in terms of the mean number of productions of of the main determiner. Participants regularize in all of the experimental conditions by over-producing the main determiner, and they do so more often in the conditions with more noise determiners.

Smith and Wonnacott (2010) investigated the regularization of plural markers. Participants were trained on various scenes that contained one of four objects (a cartooned cow, pig, giraffe, and rabbit) that were presented either as single animals or as pairs of animals (pairs were always of the same animal type). In each scene, the animals performed a "move" action, depicted with an arrow on the screen, and sentence describing the scene was presented below. Each sentence was composed of a verb (meaning "move"), noun (the animal name), and plural marker ("fip" or "tay"), in that order. When a pair of animals were shown, the sentence included one of the plural markers, but when one animal was shown, no marker was shown in that slot. Plural markers were presented randomly with a skewed, 0.75/0.25 probability, such that each participant saw one of the markers in 75% of the plural sentences and the other one in 25% of the plural sentences. In a testing phase, participants were shown a scene and prompted to type a sentence describing that scene. This experiment was conducted in an iterated learning format, such that the sentences that one participant produced served

as the training sentences for the next learner. Over five generations of learners, participants eliminated unpredictable variation in the language, but in a slightly different sense than the experiments described so far. In this experiment, participants did not regularize by eliminating one of the markers: they probability matched the plural marker frequencies, maintaining the variation in that language's plural marker forms. Instead, they regularized by forming deterministic mappings between plural markers and specific animals, such that two cows, for example, would always be marked with "fip", and two pigs, giraffes, and rabbits would always be marked with "tay". Figure 3.5 shows the results of this Smith and Wonnacott (2010). The left-hand pane shows that only a few chains regularized by eliminating one of the variants. These are the chains that level off at frequency counts of 0 or 16. Most learners maintain both marker variants in their productions. The right-hand pane shows that regularity in the co-occurance between plural markers and animal types emerges over the generations. Smith and Wonnacott (2010) use conditional entropy as a measure of regularity (further explained in Section 3.1). Lower conditional entropy indicates more regular mappings between plural markers and animals, and a conditional entropy of zero indicates fully regular (i.e. deterministic) mappings where each plural marker occurs with one animal exclusively. The main result here is a gradual drop in the average conditional entropy of the languages produced by the 10 chains, meaning that each generation of participants tends to regularize the linguistic input they receive.

Wonnacott and Newport (2005) investigated regularization of word orders in an artificial language learning task. First, participants were trained on the vocabulary items separately (5 nouns and 4 transitive verbs). Then, participants were trained on 24 sentence descriptions of different scenes. These training sentences varied in word order such that 66% of the sentences occurred in verb-object-subject (VOS) order and 33% were verb-subject-object (VSO). In the testing trials, participants were shown a scene and asked to describe it in the language they had learned. There were two testing conditions: one in which they were tested on six scenes they had seen during sentence training, and one in which they were tested on six scenes they had not seen. These novel scenes contained the vocabulary items that participants had been trained on in the vocabulary training phase, but not the sentence training phase. Participants regularized the word order of their productions when describing novel scenes: 75% of participants fully-regularized these novel scenes by producing only one word order on all of their testing trials. However, only 13% of subjects did this in the old scene condition. This study showed that participants tend to regularize word order, but

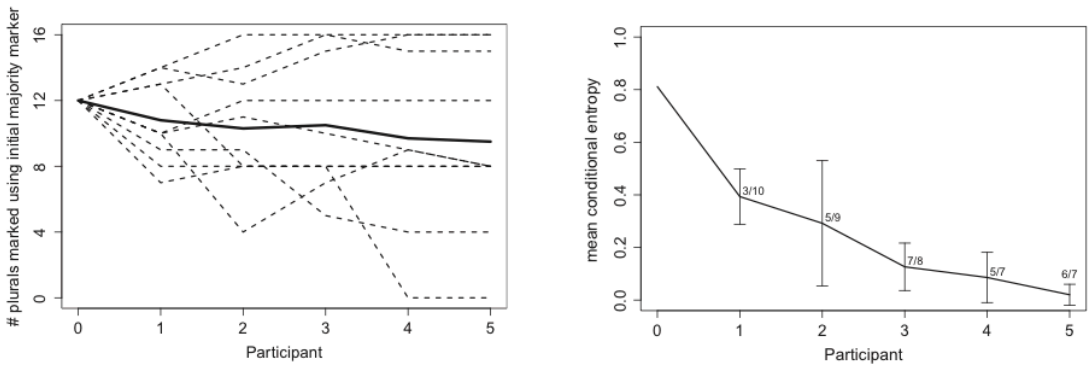


Figure 3.5: The results of the learning experiment in Smith and Wonnacott (2010). Left: the trajectories of ten iterated learning chains, through different frequency counts of the majority plural marker in the initial training set. The solid line shows the mean frequency of that plural marker. Right: the mean conditional entropy of each set of plural markers, conditioned on animal type, for all ten chains, per generation. Error bars are 95% confidence intervals. The printed fractions show the number of languages that had significantly lower conditional entropy than that expected by chance (i.e. via binomial drift of plural marker variants across animal types).

do so significantly more when using novel vocabulary items that had not been practiced in a particular word order.

The experiments above all show that adult learners tend to regularize unpredictable linguistic variation and do so for different linguistic units (nouns, determiners, plural markers, and word orders). But why does this happen? Why don't participants reproduce their training sets perfectly? And why do their "errors" in production eliminate variation rather than introduce even more variation? Humans are excellent at tracking the statistical properties of linguistic features (Fiser and Aslin, 2001; Maye et al., 2002; Newport and Aslin, 2004; Saffran et al., 1996; Saffran, 2003; Vouloumanos, 2008) as well as events in their environment (as discussed in Section 2.1 with regard to the probability matching literature: Hasher and Zacks (e.g. 1979); Gelman (e.g. 1998); Vulkan (e.g. 2000)). Given that there are a variety of instances in which adult learners do probability match, exemplified in Experiment 1, why do they regularize linguistic input? There are two main explanations for this in the literature. One is that human learners possess an innate, and domain-specific bias for regularizing linguistic input (DeGraff, 1999; Becker and Veenstra, 2003; Lumsden, 1999) which should be especially pronounced in child learners (see Bickerton (1984)'s Language Bioprogram Hypothesis). The other is that regularization is due to domain-general cognitive constraints on frequency learning, with a special focus on memory limitations (Hudson Kam and Newport, 2005, 2009; Hudson Kam and Chang, 2009).

The Less is More hypothesis (e.g. Goldowsky and Newport, 1993; Newport, 1990; Ludden and Gupta, 2000; Chin and Kersten, 2009) is a memory-based explanation for regularization behavior that may apply to children's tendencies to regularize more than adults in language learning tasks (Hudson Kam and Newport, 2005, 2009; Hudson Kam and Chang, 2009) and non-linguistic frequency learning tasks (Weir, 1964; Derks and Paclisanu, 1967; Myers, 1976). Having a limited memory capacity may cause learners to disregard or forget lower-frequency observations and thus only produce the higher-frequency alternative, resulting in regularization behavior.

In regard to this hypothesis, Hudson Kam and Chang (2009) showed that alleviating cognitive load for adult language learners led them to probability match more. Participants were taught to describe various scenes in an artificial language with variable determiners (similar to Hudson Kam and Newport (2009), described above). Then, in the testing trials, participants were prompted with a scene and produced a sentence in the artificial language to describe it. There were three testing phase conditions, one in which participants produced sentences from memory, one in which the noun from the sentences were provided for the scene, and one in which all of the words in the artificial language were provided on flash cards and participants assembled the cards to describe the scene. Participants in the first condition regularized as in Hudson Kam and Newport (2009). However, participants in the other two conditions, where word recall was aided, probability matched more among the determiner forms.

Vouloumanos (2008) also addresses the Less is More hypothesis by investigating adult learners' sensitivity to fine-grained statistical information in lexical variation consisting of homonyms: word-meaning pairings in which one word co-occurs with two or more objects. In the first experiment in this paper, participants were trained on 12 different objects in an artificial language learning task. Four of these objects were named with one word, four were named with three words (in a 8:1:1 ratio) and four were named with four words (in a 6:2:1:1 ratio). Then, in a testing phase participants were shown a word along with two objects and told to selected the object that went "best" with that word. In every test trial, one of the test objects had co-occurred more often with the word than the other during training. Here, the "correct" response would be to choose the object that had the higher co-occurrence. Participants had very low error rates in this task, regardless of the particular relative co-occurrence rate between test objects. For example, if object A occurred with the word 6 times and object B occurred with the word 1 time, participants chose object A most often. Likewise, object A occurred with the word only 2 times and object B occurred with the word 1

time, participants still chose object A most often. This means that participants are able to encode statistical information about low-frequency linguistic variants. In the second experiment, the training stimuli were manipulated to contain much higher levels of variation. Here, all 12 objects were named with four words in a (in a 6:2:1:1 ratio). The results of this experiment showed that participants maintained low error rates for the object that occurred 6 times, however the error rate on the low-frequency items was significantly higher than in Experiment 1. Therefore, increasing the variability of the stimuli, such that participants had to track more objects with in more variable mappings (6:2:1:1 vs 10:0) lead to worse encoding or recall of observed proportions of low-frequency variants. Therefore, memory limitations could be a possible driver of regularization behavior, but in a finely graded way that depends on particular properties of the distribution over observations.

This observation is corroborated by Gardner (1957) in a typical, non-linguistic frequency prediction task in which adult learners had to predict which of several lights would flash on any given trial. When participants observed two lights flashing in a 70:30 or 60:40 proportion, they probability matched in their predictions. However, when participants observed three lights flashing in a 70:15:15, 70:20:10, 60:20:20, or 60:30:10 proportion, they over-predicted the most frequent light and under-predicted the infrequent ones. This suggests that participants will regularize more when they must track more frequencies concurrently.

In a more direct investigation of the effect of working memory capacity on regularization, Kareev et al. (1997) reported an effect of individual differences in working memory capacity (as determined by a digit-span test) on participants' perception of the correlation of two probabilistic variables. Participants with lower capacity overestimated the most common variant, whereas participants with higher capacity did not. Dougherty and Hunter (2003) reported that participants with lower working memory were less likely to consider alternative choices in an eight-item prediction task and were less likely to consider the low-frequency alternatives than participants with higher working memory. However, differences in working memory capacity (either between children and adults, or as elicited by simpler and more complex stimuli sets) may not completely account for the full amount of regularization behavior elicited in artificial language learning tasks.

In a targeted investigation of linguistic frequency learning Perfors (2012) shows that placing constraints on memory encoding does not effect regularization behavior. Participants were trained on an artificial language learning task based on Hudson Kam and Newport (2009). The vocabulary consisted of 10 nouns deterministically mapped to 10 different images (i.e meanings). Nouns were al-

ways accompanied by one of 5 determiners that varied randomly in a 6:1:1:1:1 ratio. During the training phase, while participants were being taught this language, they were also presented with one of six tasks that added a particular type of cognitive load to their encoding of the language. These were 1) *verbal load*: participants read a sentence and judged whether or not it was grammatical, 2) *operational load*: participants saw a simple equation and judged whether or not the answer was correct (ex: $2/2 + 4 = 5$), 3) *low concurrent load*: participants had to remember a sequence of three letters and report them at the end of each trial, 4) *high concurrent load*: participants had to remember a sequence of six letters and report them at the end of each trial, 5) *concurrent operational load*: participants had to provide the answer to an equation, and 6) *concurrent verbal load*: participants had to memorize a sequence of four, three-letter nonsense words and report them at the end of each trial. These conditions were compared to a control condition, in which the language was learned without an additional task during training. The results show that participants did not regularize in any of the experimental conditions. In fact, the trend appeared to go in the opposite direction, with higher cognitive load leading to a more pronounced under-production of the main determiner.

These results suggest that imposing memory constraints on learners does not necessarily lead to regularization during language learning. However, it is still an open question whether cognitive load during the production phase of the task, which places constraints on memory recall, leads to regularization behavior in artificial language learning tasks. Additionally, it could also be the case that linguistic regularization is primarily driven by participants' interpretation of the goal of the task or by other pragmatic factors (Perfors, in preparation), such that what participants think they are supposed to *do* with the linguistic statistics they have encoded determines whether or not they make variable or regular productions. It is conceivable that regularization due to pragmatic factors could override the nuanced differences in regularization behavior due to memory constraints on frequency encoding or recall in linguistic frequency learning tasks.

The remaining experiments in this thesis were designed to address these issues. Experiment 2 investigates the relative contribution of domain-general (due to single versus concurrent frequency learning) and domain-specific (due to the use of linguistic and non-linguistic stimuli) drivers of regularization behavior in the basic frequency learning task. Experiment 3 takes a closer look at the levels of variability in the particular training sets and their effect on regularization behavior. Experiment 4 compares regularization behavior due to modulations of cognitive load in the observation versus production phase. Finally, Experiment

5 investigates the task framing and goal of the production task on regularization behavior. But before I present these experiments, I will explain a little more about what regularization behavior is and further develop our conceptual toolkit for understanding and describing this behavior.

3.1 An information theoretic definition

The goal of this section is to develop a definition of regularization that is explicit enough to be used for quantifying it. As we saw in the previous section, regularization is the process of eliminating variation in language, and this can happen at many different levels in a linguistic system: phonological, morphological, lexical, syntactic. More specifically, we can think of it as a process that compresses language across some dimension. The regularization experiments reviewed above all zoom in on a particular dimension of an artificial language and examine the *distribution of variation* at that level before and after learning. As for the studies targeting the morphosyntactic level, Hudson Kam and Newport (2009) and Perfors (2012) look at distributions of determiners and Smith and Wonnacott (2010) look at distributions of plural markers. At the lexical level, Vouloumanos (2008) looks at distributions of homonyms and Reali and Griffiths (2009) look at distributions of synonyms. And at the syntactic level, Wonnacott and Newport (2005) and Culbertson et al. (2012) look at distributions of words orders. Each of these studies deals with a particular set of competing linguistic *variants*, which are determiners, plural markers, nouns, or word orders.

This literature describes regularization as a process that leads one of the variants (usually the most frequent one) to be over-represented in participants' productions. This definition is fairly informal, and thus, different papers develop different measures of linguistic regularity, most of which are based on the frequency of the majority variant. Additionally, in the experiments where more than one input distribution was tested (Vouloumanos, 2008; Hudson Kam and Newport, 2009), a certain property of the distributions are found to modulate the degree to which participants regularize. Hudson Kam and Newport (2009) identify this property as the *scatter* of a distribution. For example, their training condition 2 consists of a 30%, 30%, 20%, 20% global distribution over 4 different determiners, and training condition 3 consists of a 30%, 30%, 10%, 10%, 10%, 10% global distribution over 6 different determiners. Here, condition 3 is more scattered than condition 2. Vouloumanos (2008) identifies this property as the *variability of the learning context*. For example, the variability of the learning

context is lower when objects co-occur with an 80%, 10%, 10% distribution over 3 different words than when objects co-occur with a 60%, 20%, 10%, 10% distribution over 4 different words. All of these distributions over variants, regardless of what linguistic level we are targeting, can be quantified in terms of their structure and predictability via information theoretic notions of entropy.

A few existing papers on topics closely related to regularization have utilized information-theoretic definitions of structure in language. Tamariz and Smith (2008) and Cornish et al. (2009) develop a measure of compositional structure in language called RegMap. Their latter formulation of this measure modifies the information-theoretic definition of conditional entropy into an indicator of the "degree of confidence that a signal element consistently predicts a meaning element". They apply this measure to the iterated artificial language learning experiment of Kirby et al. (2008) to verify the emergence of compositional structure over several generations of learners. Ferrer i Cancho and Solé (2003) use the conditional entropy of signal-meaning mappings in a model of Zipf's *principal of least effort* to quantify a hearer's effort in decoding a message. This formulation grounds decoding effort in the structure of the language: when signals predict meanings with greater accuracy, conditional entropy is low and hearers expend little effort in decoding the message. And as mentioned in the previous section, Smith and Wonnacott (2010) also use conditional entropy as a measure of linguistic structure and stand out as the one example of this in the regularization literature to date.

In the following sections, I will extend the use of conditional entropy (as a measure of linguistic structure) to a measure of the amount of regularization that occurs between two signal-meaning mappings in successive time steps, and explain how conditional entropy not only describes the regularity of mappings in a language (as in Tamariz & Smith, 2008; Cornish et. al, 2009; Ferer i Cancho and Solé, 2003; Smith & Wonnacott, 2010), but also describes the regularity of linguistic items, such as word orders, lexical items, and morphosyntactic markers. The latter extension could serve as a common metric for regularization in experiments such as Reali & Griffiths, 2009; Hudson Kam & Newport, 2009; Smith & Wonnacott, 2010; Wonnacott & Newport, 2005; Vouloumanos, 2008.[3]

[3]Smith and Wonnacott (2010) are cited here too because this measure would apply to their data in the left-hand pane of Figure 3.5, which they did not quantify in information-theoretic terms.

3.1.1 Shannon entropy

Entropy is a measure of the unpredictability of a random variable. Let's walk through the calculation of entropy with our example of marbles in bags. If a bag contains many marbles, of which half are blue and half are orange, and one marble is randomly drawn from this bag, anyone who tries to guess what color that marble is will do no better than chance. Here, the unpredictability of the marble color is at its maximum. This unpredictability can be quantified by the equation for Shannon entropy (Shannon, 1948)[4]:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{3.1}$$

where X is the set of marble colors (x_1 = blue and x_2 = orange). In the case of the 50/50 bag, the probability of $x_1 = x_2 = 0.5$ and the entropy of $X = -(0.5 \cdot \log_2(0.5)) + (0.5 \cdot \log_2(0.5)) = 1$ bit. This means that 1 bit of information is required to be able to say, with certainty, what the color of the marble is. More intuitively, one bit of information is equivalent to the answer of one binary question. The binary question in this case would be *is the marble blue or orange?*[5] If we change the relative proportion of marbles in the bag, then the amount of information we need to predict the color of the draw changes. In a bag of 70% blues and 30% oranges, we can be a little more sure about what color the draw will be: probably blue. Here, $x_1 = 0.7$, $x_2 = 0.3$, and $H(X) = 0.88$ bits. This means we only need to ask 0.88 of a binary question to be sure that the color was blue or orange. And if the bag contains all blue marbles and no orange marbles, we are certain the draw will be blue. Here, $x_1 = 1$, $x_2 = 0$, $H(X) = 0$ bits, and we don't have to ask any questions.

Entropy can also be calculated for non-binary variables. Suppose there are three marble colors in the bag (x_1 = blue, x_2 = orange, and x_3 = yellow). If the proportion of these colors in the bag are $x_1 = 0.5$, $x_2 = 0.25$, and $x_3 = 0.25$, then $H(X) = 1.5$ bits. This means we have to ask an average of 1.5 (optimally ordered) binary questions before we're certain what the color is. So first, we should ask *is it blue or not?* Half of the time (when the marble is blue) we will be certain of the color. But half of the time we won't and we'll have to ask a

[4]All references to entropy in this thesis will be to Shannon entropy. Other notions of entropy exist in the field of physics and I want to clarify that I will not be referring to these.

[5]If entropy is computed using a log with a base that equals the number of values the variable can take (i.e. the number of variants), then entropy is always bounded by 0 and 1. The entropy of an even distribution of blue, orange, and yellow marbles computed with $\log_3$ would be 1 trit, equivalent to asking one ternary question: *is it blue, orange, or yellow?*

further question: *is it orange or not?* Since we always have to ask the first binary question, that takes one bit. And since half of the time we have to ask the second binary question, that takes 0.5 bits (i.e. one binary question half of the time is 1 bit $\cdot \frac{1}{2} = 0.5$ bits). The total here is $1 + 0.5 = 1.5$ bits. If the questions were asked in a non-optimal order, we'd arrive at a higher number of average questions until we were certain of the color. For example, if we first ask *is it yellow or not?*, then $\frac{1}{4}$ of the time we are certain, but $\frac{3}{4}$ of the time we have to ask another question: *is it blue or not?* With this ordering, we will have to ask an average of $1 \cdot 1 + 1 \cdot \frac{3}{4} = 1.75$ binary questions. Thus, entropy is the average number of optimally ordered questions that it takes to determine the outcome of an random process with absolute certainty.[6]

So far, we have been calculating the entropy of the generating process (on the basis of the proportions of marbles in the bag), but entropy can also be calculated directly on data sets. Here, the data would be sets of draws from the bag and entropy would tell us about the structure in the *observed behavior* of this random marble generating process. For example, a set of 40 marble draws in which 20 are blue and 20 are orange (i.e. $x_1 = 0.5$, $x_2 = 0.5$) has an entropy of 1 bit. Likewise, entropy can be calculated on data sets where random variables are linguistic variants. The entropy of Hudson Kam and Newport (2009)'s distribution over 6 determiners where $x_1 = 0.3$, $x_2 = 0.3$, $x_3 = 0.1$, $x_4 = 0.1$, $x_5 = 0.1$, and $x_6 = 0.1$ is 2.37 bits. And the entropy of Vouloumanos (2008)'s distribution over 4 homonyms where $x_1 = 0.6$, $x_2 = 0.2$, $x_3 = 0.1$, and $x_4 = 0.1$ is 1.57 bits.

We know that in language, variation tends to be conditioned on other variables, such as determiners on noun classes, signals on meanings, and syntactic structures on social register. Entropy is easily extended to account for this kind of conditioned structure as well.

3.1.2 Conditional entropy

The equation for conditional entropy is as follows (Shannon, 1948):

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \tag{3.2}$$

This tells us about the entropy of a set of variants (X), weighted by the set of contexts (Y) in which they occur. Let's take an example where X is a set of words and Y is a set of meanings. In this case, conditional entropy tells us about how

[6] Thanks to Jim Crutchfield and Ryan James for explaining entropy to me via this question-based interpretation.

much information meanings carry about words in a particular linguistic system: if we know what someone's intended meaning is, the conditional entropy in bits tells us how many binary questions we need to ask to be certain what word will be said. Likewise, if we show a participant an object in an artificial language learning experiment, conditional entropy tells us how predictable the participant's usage of the language is.

Mappings of words to meanings can be deterministic as in Figure 3.6a, where meanings carry perfect information about words, and the conditional entropy is at its minimum at 0 bits (See Box 3.1 for a walkthrough of this calculation). Conversely, conditional entropy is at its maximum when meanings carry no information about words: this is the case for a fully-connected mapping, such that all words map to all meanings with equal probability. Here, maximum conditional entropy is not bounded by 1, but depends on the total number of words in the system. Mappings with intermediate values of conditional entropy are those that contain some synonyms (where one meaning can be named by two words) and/or homonyms (where one word can refer to two different meanings). Figure 3.6b gives a rather unwieldily mapping where there are two homonyms, *kal* and *buv*, and two pairs of synonyms: (*kal*, *mig*) and (*kal*, *buv*). If each meaning occurs with the probabilities 0.3, 0.5, 0.2 (from left to right) and the synonyms in each pair occur with probabilities 0.3 and 0.7, the conditional entropy of this mapping is 0.71 bits (See Box 3.1 for a walkthrough of this calculation). If instead, the synonyms occur with a 50/50 probability, this mapping is less predictable, with conditional entropy at 0.8 bits. Figure 3.6c shows a mapping that only has two pairs of synonyms and one deterministic mapping. The conditional entropy of this mapping is identical to that of Figure 3.6b. This is to illustrate the directionality of the conditional entropy calculation: both Figure 3.6b and c are identical mappings as far the conditioning of words on meanings is concerned.

However, if we flip things around and calculate the entropy of meanings conditioned on words, where X would be the set of meanings and Y would be the set of words, then homonyms become the relevant source of variation. It is important to note here that $H(X|Y) \neq H(Y|X)$. The entropy of words conditioned on meanings is not necessarily equal to the entropy of meanings conditioned on words, for the same mapping. However, $H(X) - H(X|Y) = H(Y) - H(Y|X)$ (Cover and Thomas, 1991, p.17). This means that conditional entropy quantifies the structure of a system in relation to one of the variables, while taking the entropy of that variable into account. When the entropy of the conditioning variable is at its maximum (as would be the case where all meanings (Y) occur with equal probabilities), then $H(Y) = 0$ and the conditional entropy of the sys-

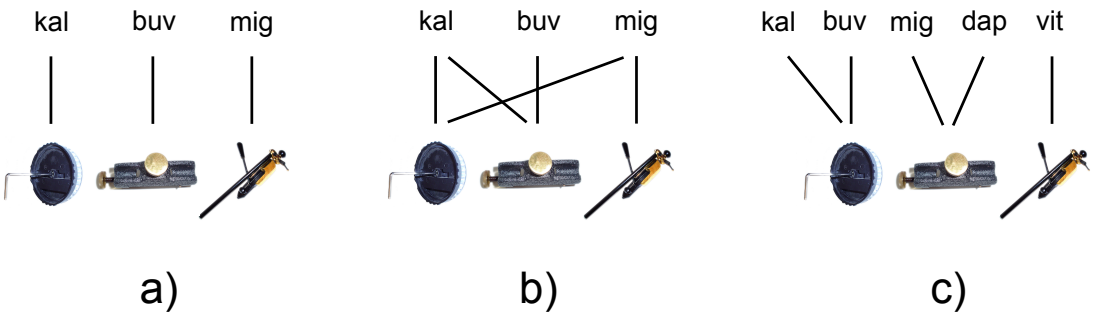


Figure 3.6: Example word-meaning mappings (referred to in Example 3.1).

Example 3.1. *Conditional entropy calculation*

The amount of structure in a set of mappings between words and meanings can be quantified by the conditional entropy of X (the set of words) given Y (the set of meanings) using equation 3.2.

Let's calculate the conditional entropy of the example mappings in Figure 3.6.

Let y_1, y_2, and y_3 be the three objects (i.e. meanings) enumerated left to right. Let $x_a, ...x_n$ be the n words in each mapping, enumerated left to right.

Let the meanings occur 30%, 50%, 20% of the time, respectively, so that $y_1 = 0.3, y_2 = 0.5, y_3 = 0.2$.

a) This mapping is deterministic.
$\{x_{kal}, x_{buv}, x_{mig}\}|y_1 = \{1, 0, 0\}$
$\{x_{kal}, x_{buv}, x_{mig}\}|y_2 = \{0, 1, 0\}$
$\{x_{kal}, x_{buv}, x_{mig}\}|y_3 = \{0, 0, 1\}$

$$\begin{aligned} & 0.3 \cdot [(1 \cdot \log_2(1)) + (0 \cdot \log_2(0)) + (0 \cdot \log_2(0))] = 0 \\ & 0.5 \cdot [(0 \cdot \log_2(0)) + (1 \cdot \log_2(1)) + (0 \cdot \log_2(0))] = 0 \\ + \; & 0.2 \cdot [(0 \cdot \log_2(0)) + (0 \cdot \log_2(0)) + (1 \cdot \log_2(1))] = 0 \\ \hline & 0 \text{ bits} \end{aligned}$$

b) Let's assume 30/70 weights on each pair of synonyms, such that y_0 is named with *kal* 30% of the time and *mig* 70% of the time.
$\{x_{kal}, x_{buv}, x_{mig}\}|y_1 = \{0.3, 0, 0.7\}$
$\{x_{kal}, x_{buv}, x_{mig}\}|y_2 = \{0.3, 0.7, 0\}$
$\{x_{kal}, x_{buv}, x_{mig}\}|y_3 = \{0, 0, 1\}$

$$\begin{aligned} & 0.3 \cdot [(0.3 \cdot \log_2(0.3)) + (0 \cdot \log_2(0)) + (0.7 \cdot \log_2(0.7))] = -0.264 \\ & 0.5 \cdot [(0.3 \cdot \log_2(0.3)) + (0.7 \cdot \log_2(0.7)) + (0 \cdot \log_2(0))] = -0.441 \\ + \; & 0.2 \cdot [(0 \cdot \log_2(0)) + (0 \cdot \log_2(0)) + (1 \cdot \log_2(1))] \quad = \quad 0 \\ \hline & 0.71 \text{ bits} \end{aligned}$$

tem is just the (unweighted) average of the entropy of words (X) per meaning. Most regularization experiments control the context stimuli so that they appear uniformly. For example, Reali and Griffiths (2009) control their object stimuli so that each object appears 10 times per participant, but the linguistic variants (the names for those objects) vary in frequency. In these cases, referring to the conditional entropy of the mapping is equivalent to referring to the (unweighted) average of entropy of each synonym set.

Now that I have established conditional entropy as a measure of a mapping's regularity, I will turn to a measure of regularization.

3.1.3 Regularization is a drop in conditional entropy

At the beginning of this chapter, regularization was described as two processes: 1) the elimination of free variation by eliminating competing variants and 2) the elimination of free variation by conditioning variants on different contexts.

Defining regularization in terms of conditional entropy accounts for both of these processes on one scale. When a system loses variants, the conditional entropy of that system goes down, and when a system gains more predictable mappings, conditional entropy also goes down. For example, if the mapping in Figure 3.6b were to lose the variant *buv*, then the conditional entropy would drop by 0.44 bits. And if the mapping of (*kal*,*mig*) were to become more predictable, by changing the weights from (0.3,0.7) to (0.2,0.8), then the conditional entropy would drop by 0.05 bits. Thus, in terms of conditional entropy, moving from (0.3,0.7) to (0.2,0.8) and increasing predictability, or from (0.3,0.7) to (0,1) and eliminating a variant, is change of the same type: an increase in regularity. This leads me to the following definition of regularization:

> **Regularization** is a drop in entropy of variants (X) when conditioned on certain contexts (Y) over time.

This is a definition of regularization in terms of an outcome (human behavior), not process (the mechanisms that create regularization behavior). I have taken this approach because it is a useful framing for experimentalists, who largely use participant behavior to infer the cognitive processes behind regularization. Over the course of this thesis, we will see that many different processes can elicit regularization behavior, such as cognitive load during frequency learning (Chapter 3), domain-specific biases for linguistic stimuli (Chapter 3), strategies

for solving a coordination game (Chapter 4), and even cultural drift (Chapter 5). Having a universal metric for quantifying regularization behavior has several benefits. First, it allows experimentalists to discuss the different cognitive processes that give rise to regularization behavior in the same, comparable terms. Second, when several regularization processes are at play contemporaneously, it allows the relative contribution of each process to be compared. Third, it should make model selection over alternative hypotheses more comprehensive because all processes can be formulated in terms of their effect on conditional entropy over time. Fourth, it allows regularization at different levels of language to be compared. The existing literature discusses regularization as if it were one monolithic process, with experimental examples at phonological, morphological, lexical, and syntactic levels. It is quite possible that different processes give rise to regularization at these different levels. Although we should not conflate these processes, we should maintain a universal definition of regularization behavior so that differences in behavior, due to these different processes, are more readily visible. And last, in relation to all of the previous points, a universal measure of regularization allows different experiments to be compared. As mentioned earlier, Hudson Kam and Newport (2009) and Vouloumanos (2008) both report that the entropy of training distributions modulate participant regularization behavior, and Hudson Kam and Newport (2009) explicitly shows that higher-entropy distributions are regularized more. What are the specific drops in entropy here? Are they the same across experiments, or do they differ in systematic ways? If there's a difference, is this due to different processes underlying the regularization of determiners and homonyms? Or is this difference just because higher-entropy distributions *can* be regularized more and the two experiments used training sets that systematically differed in entropy? Being able to answer questions such as these will likely sharpen the focus of the next generation of experimental design in regularization research.

Most language learning experiments control their stimuli such that all of the meanings occur the same number of times. This constitutes a uniform distribution over contexts (Y) where no context is more likely than any other, so $H(Y) = 0$. When this is the case, the conditional entropy, $H(X|Y)$, is equivalent to an unweighted average of $H(X)$ per y (refer back to Equation 3.2). All of the experiments in this thesis present participants with uniform distributions over contexts. In some of my analyses of participants behavior, I will refer to the conditional entropy score as the drop in the average entropy of all synonym sets over time. Please keep in mind that these two framings are identical.

Before leaving this section, I must point out one thing that seems to be over-

looked in the regularization literature. On the basis of this regularization definition, it seems intuitive that *unbiased probability matching* behavior should entail no drop in conditional entropy, however the specific amount of error on unbiased probability matching leads to a small amount of regularization. This is (partially) why drifting populations eventually end up losing variation over time. Only *perfect probability matching*, where there is no error, leads to no drop in entropy over time. This may seem counter-intuitive, but referring back to the binomial drift distributions in Figure 1.2, it is clear that error can lead to transitions between input and output ratios that are regular (ex: 5:5 $\rightarrow$ 7:3), but also to transitions that increase variability (ex: 7:3 $\rightarrow$ 5:5). Depending on the initial distribution of input ratios, different numbers of regularity-increasing or variability-increasing transitions will be produced by drift. If we start with a uniform distribution over input ratios (as was the case in Experiment 1, where 32 participants were trained on each input ratio: 5:5, 6:4, 7:3, 8:2, 9:1, and 10:0), then applying one generation of $N = 10$ binomial drift yields an average change in entropy of -0.05 bits. This means that the resulting distribution of output ratios, depicted in in Figure 1.2 differs from their respective input ratio by -0.05 bits on average. This was calculated by subtracting the entropy of each output ratio from the entropy of its input ratio, then weighting the resulting values by their probability under one generation of drift (i.e. the probabilities shown in Figure 1.2), and averaging them. All unbiased sampling processes that contain some amount of error should yield a small amount of regularization *behavior* and it's the amount of variance in this sampling error that determines how much regularization behavior occurs.

So the above example proves that an unbiased sampling process, which by definition is not a regularization process, can produce regularization behavior. This issue suggests that experimentalists must pay special attention to how they determine whether some observed regularization behavior was the result of a regularization process or not. A good way to go about making this distinction would be 1) to state that regularization *behavior* is supported empirically when the change in entropy is significantly lower than zero and 2) to state that a regularizing *process* (such as an inductive bias for variation elimination) is supported empirically when the change in entropy is significantly lower than the change in entropy that *unbiased probability matching* would yield for that task.

This makes life difficult for experimentalists who study regularization, because it means that we must show a significant difference from a cognitive drift process in order to conclude the existence of regularization biases. A baseline can be calculated for any set of stimuli via random sampling methods like Monte-Carlo Markov Chain methods (e.g. Smith and Wonnacott, 2010), however this will de-

fine a baseline with binomial/multinomial variance. In the previous chapter, I showed that the cognitive basis of drift is not necessarily binomial. However, if the variance associated with human probability matching behavior is always lower than that of its binomial/multinomial counterpart, then change in entropy of these processes will always be lower than that of binomial/multinomial drift, and therefore, the standard random sampling methods will represent a conservative, worse-case scenario null hypothesis. For example, the restricted-variance probability matching behavior in Experiment 1 has a change in entropy of 0.027 bits, which is closer to zero than binomial drift (although it happens to have produced a little more variability rather than regularity, this was not significantly different from zero). This is good news. However, very little is known about the variance associated with human probability matching behavior and it can be impossible to determine this baseline for a particular experiment that consistently elicits regularization behavior, because there may be nothing one can do to get participants to probability match in a similar enough experiment. This point will be made concrete when I operationalize this baseline in Experiment 2.

3.2 Experiment 2: regularization biases in frequency learning

At the beginning of this chapter, I reviewed a body of research that shows regularization to be a complex phenomenon with many potential sources, both domain-specific and domain-general. Experiment 2 is designed to tease apart the domain-general and domain-specific drivers of regularization at the lexical level in adult learners, and evaluate the relative contribution of each of these sources to our overall regularization bias for language.

This experiment builds upon Experiment 1 by showing how two manipulations can take learners away from probability matching and elicit regularization behavior. The first manipulation recasts the one-item frequency learning task in Experiment 1 as a multiple-item task in which participants learn about marble draws from several different containers concurrently. This provides a picture of regularization due to the general cognitive mechanisms associated with tracking and producing multiple frequencies concurrently and relates to the work of Gardner (1957), Newport (1990), Saffran (2003), McElreath et al. (2008), Hudson Kam and Chang (2009), and Perfors (2012).

The second manipulation recasts Experiment 1 in a linguistic domain, where participants learn about one object being named with two synonyms. This al-

lows an assessment of how linguistic stimuli trigger domain-specific regularization biases during frequency learning, relating to the proposals of language-specific regularization biases discussed by Bickerton (1984), DeGraaff (1999), Lumsden (1999), and Becker and Veenstra (2003).

This yields a 2-by-2 experimental design where domain and concurrency are manipulated (Figure 3.7):

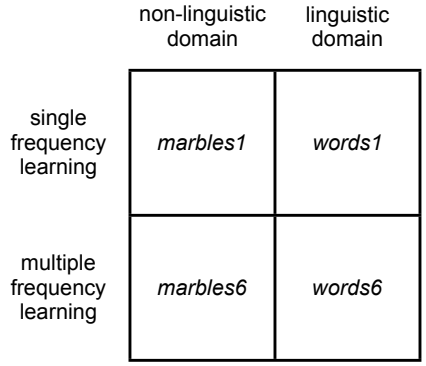


Figure 3.7: Names of the four conditions in Experiment 2.

Of these four conditions, *marbles1* is a replication of Experiment 1, with a few minor adjustments to the procedure that allow for better comparability to the three other conditions. The condition *words6* combines the two manipulations to study the full regularizing effect of the concurrent frequency learning of multiple linguistic items. This condition doubles as a replication of the artificial word learning experiment in Reali and Griffiths (2009).

We will see that both of these manipulations modulate regularization behavior. In brief, the results of this experiment show that both linguistic domain and multiple frequency learning elicit regularization behavior, whereas non-linguistic, single frequency learning elicits probability matching behavior. Thus, lexical regularization is rooted in domain-general and domain-specific sources. Furthermore, we will see that these sources independently contribute to participants' full linguistic regularization bias, as elicited in the *words6* condition.

3.2.1 Method

Participants

573 participants were recruited via Amazon's Mechanical Turk crowdsourcing platform and completed our experiment online. Participant location was restricted to the United States of America and verified by a post-hoc check of participant IP address location. 61 participants were excluded on the basis of the following criteria: failing an Ishihara color vision test[7] (15), self-reporting the use of a pen or pencil during the task[8] (10), not reporting their sex or age (6), self-reporting an age below 18 (1), or having previously participated in this or any of my experiments, as determined by their user ID with MTurk (26). More participants were recruited than necessary with the expectation that some would be excluded by these criteria. Once the predetermined number of participants per condition was met, the last participants were excluded, totaling 3 participants across all conditions. All excluded participants received the full monetary reward for the task, which was 0.10 USD in the one-item conditions (*marbles1* and *words1*) and 0.60 USD in the multiple-item conditions (*marbles6* and *words6*). The average time taken to complete the one-item conditions was 3 minutes and 46 seconds, with a standard deviation of 1 minutes and 27 seconds. The average time taken to complete the multiple-item conditions was 11 minutes and 21 seconds, with a standard deviation of 2 minutes and 6 seconds. Of the final 512 participants, 52% are female and the mean age is 33.4 (min = 18, max = 72) with a standard deviation of 11.2 years. The breakdown of participants in each of the four conditions (further described in section 3.2.1) is as follows:

Conditions:	all	*marbles1*	*words1*	*marbles6*	*words6*
Participant count	512	192	192	64	64
Age: mean	33.4	32.8	32.2	35	33.6
minimum	18	18	18	18	18
maximim	72	72	68	65	64
standard deviation	11.2	10.6	10.9	12	11.2
Sex (% female)	52%	47%	63%	47%	50%

[7]Only participants in the non-linguistic conditions were given the color vision test. I used two plates from the Ishihara color vision test: plate 4 which tests for red-green color deficiency, and plate 23 which tests for protanopia and deuteranopia (see Appendix A.2.1). Participants were excluded if they gave an incorrect answer for one or both of these plates.

[8]In an exit questionnaire.

Materials

The experiment was coded up as a Java applet that ran in the participant's web browser in a 600x800-pixel field. Photographs of 6 different containers (a bucket, bowl, jar, basket, box, and pouch) and graphically generated images of marbles in 12 different colors (blue, orange, red, teal, pink, olive, lime, purple, black, yellow, grey, and brown) served as non-linguistic stimuli (Figure 3.8). Modified photographs of 6 different novel objects (resembling mechanical gadgets) and 12 different nonsense words (buv, kal, dap, mig, pon, fud, vit, lem, seb, nuk, gos, tef) served as linguistic stimuli (Figure 3.9).

Figure 3.8: Stimuli used in the non-linguistic conditions.

Figure 3.9: Stimuli used in the linguistic conditions.

Marbles and words were organized into fixed pairs, so that certain features could be controlled for to maximize distinctiveness between the stimuli in the pair. Marble colors on the same row in Figure 3.8 constitute a pair (for example, the blue and orange marbles are paired). Likewise, words on the same row in Figure 3.9 constitute a pair (for example, *buv* and *kal* are paired).

Marble colors were chosen to differ in hue and brightness. Within-pair hue differences were greater than 120° (meaning that they were chosen from approximately opposite sides of the color wheel) and within-pair brightness differences were greater than 20%. These criteria rule out, for instance, pairings such as red and orange, red and pink, or black and grey. The RGB values for the colors used are as follows:

	blue	orange	red	teal	pink	olive	lime	purple	black	yellow	grey	brown
R	0	255	204	0	255	0	67	153	0	255	153	102
G	51	128	0	204	153	65	215	0	0	204	153	51
B	204	0	0	255	153	1	9	153	0	0	153	0

Words were also paired to be contrastive. Words within a pair differed in place of articulation and the letters they contained. Some examples of word pairings ruled out by these criteria would be *buv* and *pon* (because *b* and *p* share the same, bilabial, place of articulation) or *seb* and *tef* (because they share a letter, *e*). These stimuli are closely based on the word stimuli used in Reali and Griffiths (2009), but modified so that no words would be orthographically English, or have likely pronunciations in English. Words were presented visually and were not accompanied by auditory stimuli.

Conditions and Design

The relative contributions of domain and concurrency to linguistic regularization were investigated in a two by two experimental design, consisting of four conditions:

1) *Non-linguistic single frequency learning (marbles1)*
Participants observed two marble colors being drawn from one container at a particular ratio (for example, 5 blue marbles and 5 orange marbles drawn in random order). Participants were then asked to demonstrate what another several draws from the same container are likely to look like. They were not asked to predict specific future draws and thus no feedback was given. Participants observed 10 marble draws and produced 10 marble draws. Each participant observed one of six possible ratios: 0:10, 1:9, 2:8, 3:7, 4:6, 5:5. (To be clear, these ratios were not used to probabilistically generate the draws participants observed; these are the ratios of the actual marbles participants observed.) Each ratio was observed by 32 participants, totaling 192 participants for this non-linguistic single frequency learning condition. Container stimuli were randomized across participants: each participant only saw one container of the six in Figure 3.8. Equal numbers of participants saw each container. Marble pairs were also randomized across participants: each participant only saw one of the six marble pairs in Figure 3.8. Equal numbers of participants saw each marble pair. The full details about the observation and production regimes can be found in the next section, *Procedure*.[9]

2) *Non-linguistic multiple frequency learning (marbles6)*
This condition is similar to the *marbles1* condition, with the difference that participants observed 10 draws each from 6 different containers. Next, participants were asked to demonstrate what another several draws from the same six containers are likely to look like. Participants observed 60 draws and produced 60 draws. Containers, marble pairs, and observation ratios were randomly assigned to one another, without replacement. This means that each participant saw all six of the containers, all six of the marble pairs, and all six of the ratios, in some

[9]This condition is a replication of Experiment 1 with some modifications so that it would be comparable to the designs of the 3 other conditions in Experiment 2. The modifications are: 1) The participant sees one of six containers (not just the bag) and one of six color pairs (not just blue and orange). 2) Each production trial includes an OK button that appears after the participant has clicked on one of the marbles/words. Participants can change their choice before pressing OK. The OK button serves to recenter the mouse between production trials. There was no recenter feature in Experiment 1. 3) The test-side of the majority marble is random between trials per participant (as opposed to held constant).

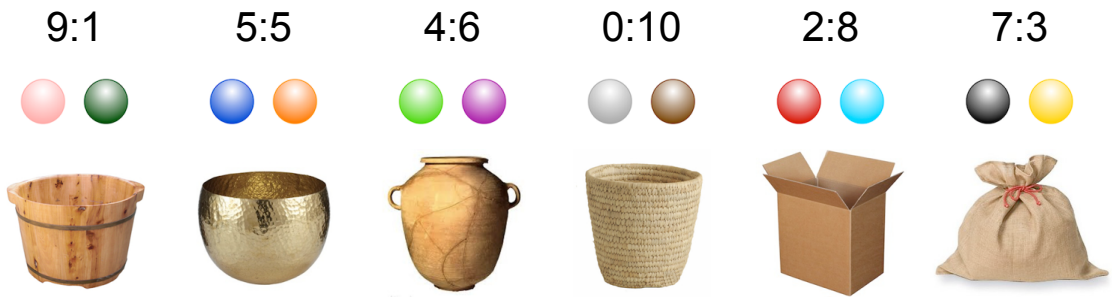


Figure 3.10: An example stimuli set for one participant in the non-linguistic multiple frequency learning condition. This participant would observe a 1:9 ratio of the pink and olive marbles drawn from the bucket, a 5:5 ratio of the blue and orange marbles drawn from the bowl, and so on. The mapping between containers, marble pairs, and ratios was randomized across participants.

combination. An example of one such assignment can be seen in Figure 3.10. These assignments were re-randomized per participant, so that no two participants saw the same containers paired with the same marbles at the same ratios. Assignments were pre-generated and screened to make sure that there was no correlation between particular containers, marble pairs, or ratios across participants. All Pearson correlation coefficients were between 0.083 and -0.064. There were 64 participants in this condition, yielding data for 384 (6x64) observed ratios.

3) *Linguistic single frequency learning (words1)*
This condition is similar to the *marbles1* condition, differing only by the use of linguistic stimuli (objects and words) instead of the non-linguistic stimuli (containers and marbles). Participants observed one object being named with two words at a particular ratio (for example, *buv* 5 times and *kal* 5 times, in random order). Participants were then asked to name the object like they had observed it being named. They were not asked to predict specific future naming events and thus no feedback was given. Participants observed 10 namings and produced 10 namings. Each participant observed one of six possible ratios: 0:10, 1:9, 2:8, 3:7, 4:6, 5:5. Each ratio was observed by 32 participants, totaling 192 participants for this linguistic single frequency learning condition. Object stimuli were randomized across participants: each participant only saw one object of the six in Figure 3.9. Equal numbers of participants saw each object. Word pairs were also randomized across participants: each participant only saw one of the six word pairs in Figure 3.9. Equal numbers of participants saw each word pair.

4) *Linguistic multiple frequency learning (words6)*

This condition is similar to the *marbles6* condition, differing only by the use of linguistic stimuli. This condition constitutes a replication of the word learning experiment in Reali and Griffiths (2009), but with different object stimuli, modified word stimuli, and participants who completed the experiment online, rather than in the laboratory. Participants observed 60 namings and then produced 60 namings. Object, word pair, and ratio assignments were randomized as in marbles6. These assignments were also pre-generated and checked to make sure that there was no correlation between particular objects, word pairs, or ratios across participants. All Pearson correlation coefficients were between 0.082 and -0.057. There were a total of 64 participants in this condition, yielding data for 384 (6x64) observed ratios.

Procedure

The experiment consisted of an observation phase and a production phase (Figure 3.11). In each observation trial, a container/object was displayed on its own for 1 second and then a marble/word was displayed above it for 2 seconds, with no break between trials. There were 10 observation trials per container/object. In each production trial, a container/object was displayed and the two marbles/words that appeared with it during observation were displayed below. This part had no time constraint. When participants clicked on one of the marbles/words, an OK button appeared between the two choices. Participants could change their choice, but when they clicked on the OK button, their current selection was registered as their answer and this choice was displayed above the container/object for 2 seconds. This OK button also served to center the cursor between trials. Production trials repeated until 10 responses were collected per container or object. Participants were not told how many observation or production trials there would be. In the single frequency learning condition, participants received a total of 10 observation and 10 production trials. In the multiple frequency learning condition, where participants observed 10 draws each from 6 containers, there were a total of 60 observation and 60 production trials. Here, the order in which the different containers appeared was randomized across observation and production trials. For example, in observation trial 1 the participant sees one draw from the box, in trial 2 they see a draw from the jar, in trial 3 they see one from the basket, in trial 4 they see one from the box again, and so on. Likewise in the production trial, the participant will be prompted to demonstrate a draw from the pouch, then the basket, then the box, and so on. The production-phase locations of the two marbles/words were randomized per trial. For example, if the two test choices were blue and orange marbles, then on some trials blue would be on the left and on other trials, it would be on the right. The marble/word in the pair that would be the more frequent item was randomly chosen per participant. For example, in a 2:8 ratio, some participants would see the blue marble 8 times, whereas others would see the orange marble 8 times.

After the production phase, participants were asked to estimate the generating ratio that underlies each observation ratio they saw. Participants chose their response with a discrete slider over 11 options of relative percentages. These 11 options ranged from {100:0, 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, 10:90, 0:100}. See Appendix A.2 for the instructions and screen shots of these sliders per condition: *marbles1* (Figure A.2), *words1* (Figure A.3), *marbles6* (Figure A.4), and *words6* (Figure A.5).

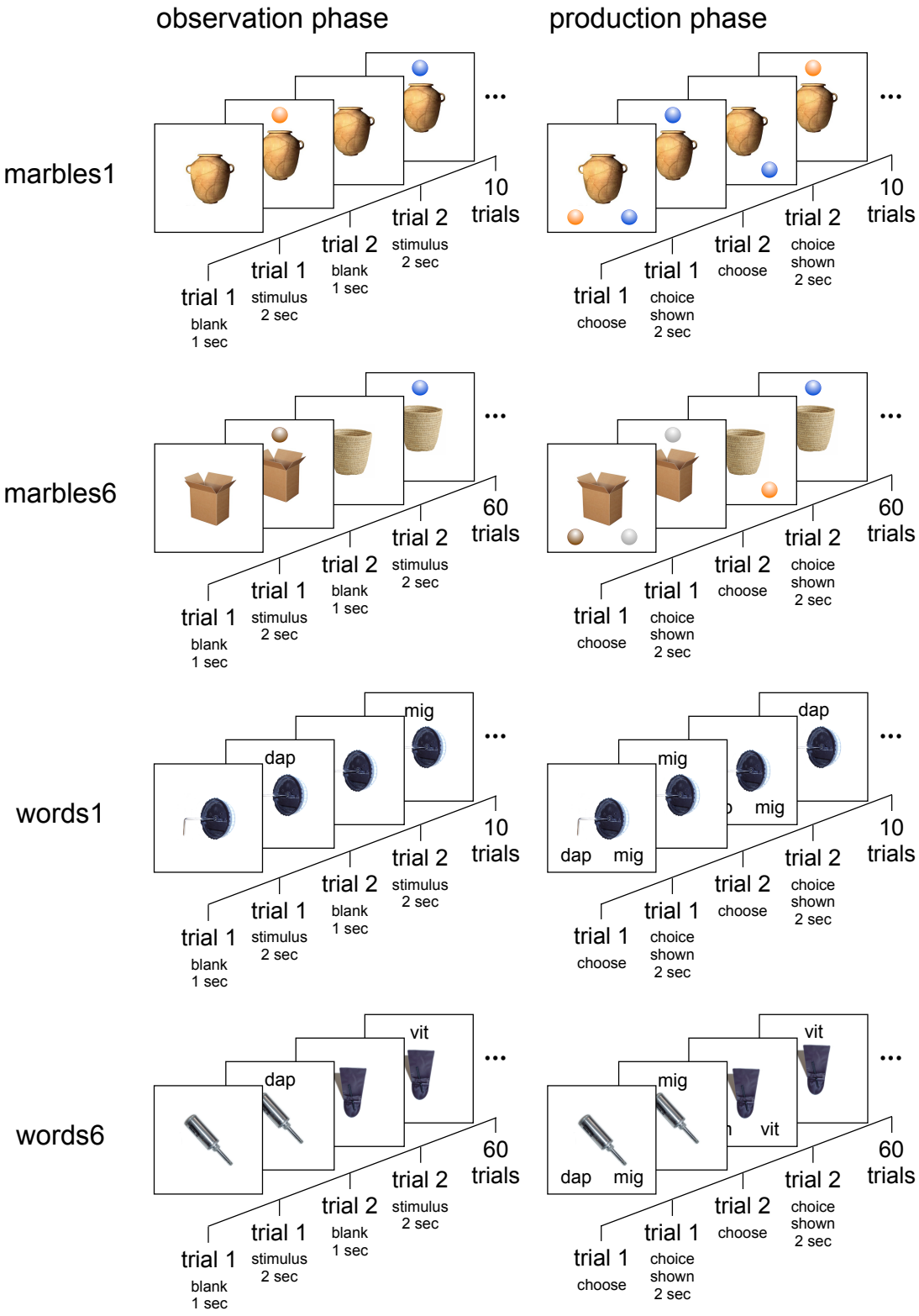


Figure 3.11: Schemas of the observation and production phases for each of the four main experimental conditions.

Next, in the one-item conditions only, participants were asked to report 1) the number of times they think they saw each marble/word in the observation phase and 2) the number of times they think they produced each marble/word in the production phase. Participants chose their response with a discrete slider over 11 options of relative counts. These 11 options ranged from {10:0, 9:1, 8:2, 7:3, 6:4, 5:5, 4:6, 3:7, 2:8, 1:9, 0:10}. See Appendix A.2 for the instructions and screen shots of these sliders: *marbles1* A.6 and *words1* A.7.

Last, participants were given seven brief exit questions. Complete instructions for all parts of this procedure and exit questions per condition can be found in Appendix A.2.

3.2.2 Results

Summary of the ratios produced

In this section, I will present the raw ratios that participants produced. Then, in the following sections, we will convert these data into the information theoretic measure of regularization (see section 3.1) and analyze the effects of the experimental manipulations on participant regularization behavior.

Figure 3.12 shows the results of Experiment 2. A visual inspection of all the results for *marbles1* suggests that participants probability matched in this condition, because the mode of each response distribution is located on the observation ratio. This falls under the behavioral definition of probability matching (see section 2.1), where participants reproduce the variability they observe, with some error. This condition seems to give a slightly noisier picture of probability matching than in Experiment 1, which it was supposed to replicate. This is most likely due to the within-participant randomization of test-side location across production trials. This means that on any given production trial, *variant x* appeared on the left- or right-hand side of the screen randomly. In Experiment 1, *variant x* always appeared on the same side across all production trials (and this was counterbalanced across participants). However, this test-side randomization does not seem to have made things too difficult for participants because most of them were still able to probability match, and consistently select *variant x* in response to the 10:0 observation ratio in all four conditions.

In the three remaining experimental conditions, participants do not appear to be probability matching. Instead, they tend to produce many fully-regular responses in the ratios 10:0 or 0:10. For example, in the *marbles6* condition for observation ratio 8:2, we see less probability matching behavior (only 17% of participants reproduced the 8:2 ratio) and more regularization behavior (36% of

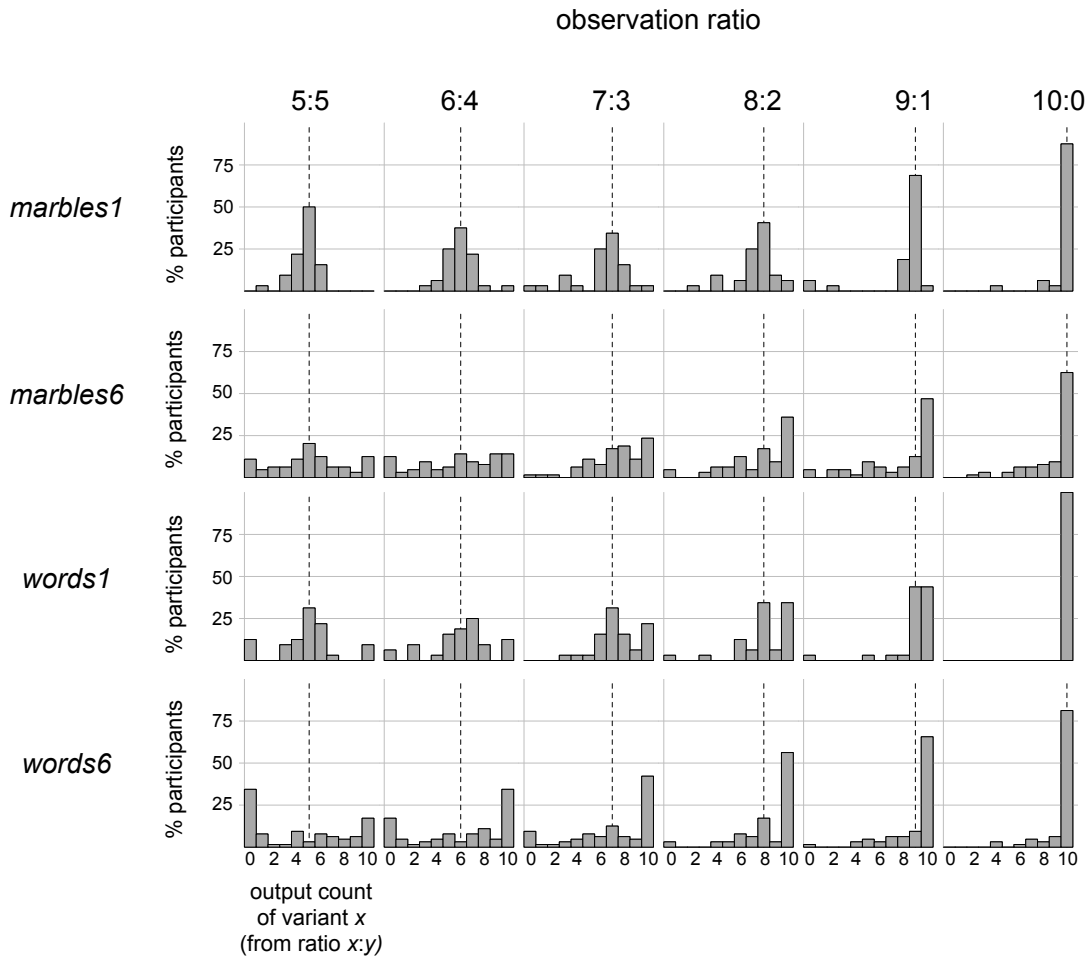


Figure 3.12: Results of the four main conditions in Experiment 2. Each row shows the results of one experimental condition (marbles1, marbles6, words1, words6). Each column corresponds to one of the six observation ratios, ranging from 5:5 (on the left) to 10:0 (on the right). Each pane contains the distribution of ratios that participants produced in response to one observation ratio. These production ratios are displayed on the x-axis as the number of times a participant produced *variant x* from the observation ratio *x:y*. *Variant x* corresponds to whatever marble/word was in the majority during the observation phase. (In the 5:5 observation ratio there is no majority variant, so a random marble/color was coded as *variant x*.) All observation ratios are indicated by a dashed line. For example, the top left panel gives the results for the 32 participants in condition *marbles1* who observed a 5:5 ratio of marbles. Here, we see that 50% of these participants also produced a 5:5 ratio, 22% produced a 4:6 ratio, 16% produced a 6:4 ratio, and no participants produced only one marble color (i.e. a 10:0 or 0:10 ratio).

participants produced the 10:0 ratio and 5% produced the 0:10 ratio). Across all observation ratios in *marbles6*, *words1*, and *words6*, some degree of full regularization is obtained. The *words1* condition seems to elicit equal amounts of probability matching and regularization, given the bimodal distributions with peaks on the observation ratios and 10:0 ratios, whereas *marbles6* and *words6* elicit more regularization behavior, with *words6* eliciting the most.

Participants tend to regularize by over-producing the majority variant. All bars on the right-most side of the panels show the participants who fully-regularized by only producing the majority variant, whereas all left-most bars show the participants who did the opposite and regularized by over-producing the minority variant. For all 5:5 observation ratios there is no minority or majority variant, but a large amount of full regularization is still obtained. This means that regularization is not completely dependent on the presence of a majority variant, but is elicited by some more general properties of the domain and concurrency manipulations in this experiment.

Regularization per condition

Now that that I have described the regularization profiles for this experiment, the first question is: do these experimental conditions elicit a significant amount of regularization? To address this question, the data in Figure 3.12 were converted into the regularization measure set forth in Section 3.1, in which regularization is defined as a drop in entropy within a pair of observation and production ratios. Change in entropy was calculated for each pair of observation and production ratios by subtracting the entropy of the observation ratio from the entropy of the production ratio. Figure 3.13 plots the average change in entropy per experimental condition. When the average change in entropy is significantly below zero, this is evidence that participants were regularizing their productions. When average change is not significantly different from zero, we can assume participants were probability matching (but refer back to Section 3.1.3 for a discussion of this point).

Significant differences from zero were assessed by using R (R Core Team, 2013) and *lme4* (Bates et al., 2013) to perform a linear mixed effects regression analysis. A model was constructed with entropy drop as the dependent variable and condition as the independent variable (i.e. fixed effect). Participant was entered as a random effect because in conditions *marbles6* and *words6*, more than one data point was gathered per participant. This violates the independence assumption of many simpler statistics, such as the t-test, and therefore these

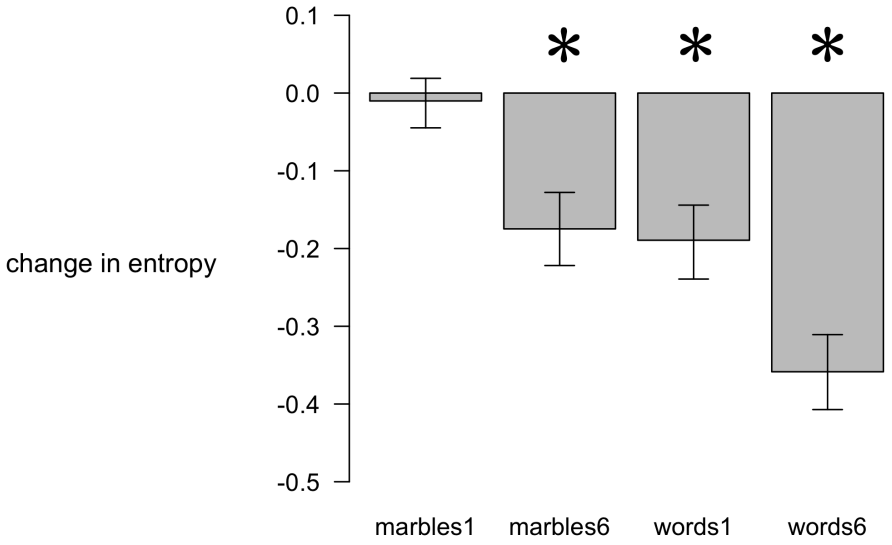


Figure 3.13: Average change in entropy within observation and production ratio pairs, per condition. Stars indicate significant difference from zero (all to $p < .001$). Error bars indicate the 95% confidence intervals computed with the bootstrap percentile method (Efron, 1979). A significant drop in entropy means that participants regularized in that condition. Non-significant differences from zero mean that participants probability matched. The bounds on entropy change for this experimental setup range from -4.04 to 1.96.

marbles1	*marbles6*	$t(1152) = -3.833, p < .001$
	words1	$t(1152) = -4.362, p < .001$
	words6	$t(1152) = -8.113, p < .001$
marbles6	*words1*	$t(1152) = -0.340, p = 0.73$
	words6	$t(1152) = -4.110, p < .001$
words1	*words6*	$t(1152) = -3.940, p < .001$

Table 3.1: Significance of pairwise comparisons between the four conditions in Experiment 2.

could not be used. No obvious deviations from normality or homoscedasticity were apparent from a visual inspection of residual plots, meaning that these data do not violate the basic assumptions of a linear mixed effects regression analysis.

The model was releveled and run four times to obtain the intercept value for each condition. Here, the intercept corresponds exactly to the mean entropy change of the condition and the regression analysis provides a t-statistic to evaluate whether or not this mean is significantly different from zero. Three of the four experimental conditions elicited a significant amount of regularization behavior: *marbles6*, $t(1152) = -5.526, p < .001$; *words1*, $t(1152) = -6.519, p < .001$; and *words6*, $t(1152) = -11.338, p < .001$.[10] In *marbles1*, change in entropy was not significantly different from zero, $t(1152) = -0.350, p = 0.73$, which confirms that

[10]all p-values presented in this chapter are for 2-tailed tests.

participants in this condition were probability matching.

Seeing that three of the four condition elicit regularization, the next logical question is: are there significant differences in the amount of regularization between conditions? The same regression model was used to address this question, because the releveled model results also provide t-scores for evaluating significant differences between conditions. Table 3.1 reports all pairwise comparisons between conditions. All comparisons, except for that between *marbles6* and *words1*, are significantly different. This suggests that both manipulations, domain and concurrency, elicit similar amounts of regularization behavior.[11] The significant difference between the probability matching condition *marbles1* and the three regularizing conditions *marbles6*, *words1*, and *words6* is also an important finding. As discussed in Chapter 2, probability matching with error can produce instances of regularization behavior and thus, is the appropriate null hypothesis to reject before claiming that a regularizing process is at play. Although these three conditions were significantly different from zero, the fact they they elicit significantly more regularization behavior than in a similar task where participants probability match provides stronger support that participants are truly regularizing in these conditions (refer back to Section 3.1.3 for a discussion of this point).

In the following section, these differences per condition will be addressed in a way that better illuminates the cognitive drivers of regularization, by looking at the effects and interactions between the two main manipulations in this experiment: domain and concurrency.

Effect of the experimental manipulations

In this section, I address the experimental manipulations that were hypothesized to modulate regularization behavior, domain and concurrency, and investigate how they are affected by the different observation ratios. To investigate the relationship between these manipulations and regularization behavior, I performed another linear mixed effects regression analysis. Entropy change was entered as the dependent variable. Domain, concurrency, and the entropy of the observation ratio were entered as the fixed effects, all with interaction terms. Participant was entered as a random effect as random intercepts. No obvious deviations from normality or homoscedasticity were apparent from a visual inspection of residual plots. Table 3.2 summarizes this full model.

[11] At least for this task where participants learn about 6 items concurrently. It is quite possible that learning about 3 or 12 items in the marbles domain will elicit significantly different amounts of regularization compared to *words1*.

model	dependent variable	fixed effects	random effects	
full model	change in entropy	domain concurrency observation ratio entropy (interactions among all)	intercepts: slopes:	participant none

Table 3.2: The full model used in the linear mixed effect regression analysis of regularization behavior in Experiment 2.

model	dom, con	con, obs	dom, obs	dom, con, obs
full model	interact	interact	interact	interact
A	interact	interact	interact	add
B	add	interact	interact	add
C	interact	add	interact	add
D	interact	interact	add	add
E	interact	add	add	add
F	add	interact	add	add
G	add	add	interact	add
H	add	add	add	add

Table 3.3: The nine logically possible relationships between the three fixed effects: domain (dom), concurrency (con), and observation ratio entropy (obs).

P-values were obtained by likelihood ratio tests, performed by an ANOVA, on the full model with the effect in question against a reduced model that omits the effect in question. If the full model is significantly better at describing the data than the reduced model, that means the effect in question is a significant predictor of entropy change. The χ^2 test statistic and p-value of the full to reduced model comparison is reported for the effect in question. (e.g. Winter, 2013).

There is a significant effect of domain $\chi^2(4) = 46.048, p < .001$, concurrency $\chi^2(4) = 105.07, p < .001$, and observation ratio $\chi^2(4) = 520.23, p < .001$ on entropy change. The directionality and meaning of these effects will be discussed after the significance of their interactions is determined.

Table 3.3 shows the nine possible relationships between variables in terms of the interactive and additive effects they can have on one another. There are several ways to assess the existence of interactions among these variables. First, a model selection procedure can be used to determine the best-fit model. Second, significance of each variable can be determined by removing, in turn, each interaction from the full model, as in models A, B, C, and D. It should be noted here that removing any one interaction necessarily removes the three-way interaction. (So A should be compared to the full model, then B, C, and D should be compared to A.) Third, significance of each variable can be determined by adding each interaction to the fully reduced model H, as in models E, F, and G. I will present each of these three assessment strategies, first to check that they

corroborate one another, and second, to use as much information as possible for inferring the presence of absence of particular interactions in the data.

Model selection. An ANOVA was run on all nine models and returned three models that were significantly better than all others. These are model B ($\chi^2(1) = 74.9409, p < .001$, loglikelihood = -278.71), D ($\chi^2(0) = 70.2497, p < .001$, loglikelihood = -281.06), and F ($\chi^2(0) = 74.6893, p < .001$, loglikelihood = -281.06). All of these models include an interaction between concurrency and observation ratio, but differential evidence in support of the interaction of domain with the other two variables: B says domain does not interact with concurrency, D says domain does not interact with observation ratio, and F says that domain does not interact with either of these variables. Recent developments in the literature on model selection techniques advocate for multi-model inference (e.g. Burnham and Anderson, 2002; Mars et al., 2012; McElreath et al., 2005), which takes into account the information provided by the best *set* of models when drawing conclusions about one's data. The conventional alternative would be to only make inferences based on the model with the maximum likelihood, which in this case is model B ($\chi^2(1) = 4.6918, p = 0.03$, from an ANOVA on models B, D, and F only). However, given that model B has a marginal p-value and only slight improvement in log likelihood, I will base my conclusions on the three best models. Together, these models provide 1) strong support that there is no interaction between domain and concurrency, 2) strong support that there is an interaction between concurrency and observation ratio, and 3) weak support that there is an interaction between domain and observation ratio.

Selective omission. First, model A was compared to the full model with an ANOVA and no significant effect of a three-way interaction was found ($\chi^2(1) = 0, p = 1$). Then, a comparison of models B, C, and D to A, found no significant interaction between domain and concurrency ($\chi^2(3) = 0.0065, p = .99$), a significant interaction of concurrency and observation ratio ($\chi^2(3) = 74.942, p < .001$), and no significant effect of domain and observation ratio ($\chi^2(3) = 4.6918, p = 0.20$). These results corroborate the model selection results above, and strengthen the case that domain does not interact with the other variables.

Selective inclusion. Models E, F, and G were compared to the fully reduced model H with an ANOVA and found no significant interaction between domain and concurrency ($\chi^2(1) = 0.0059, p = 0.94$), a significant interaction of concurrency and observation ratio ($\chi^2(1) = 74.695, p < .001$), and a significant interaction of domain and observation ratio ($\chi^2(1) = 4.4462, p = 0.03$). These results also corroborate the model selection results, but strengthen the case that domain interacts with observation ratio.

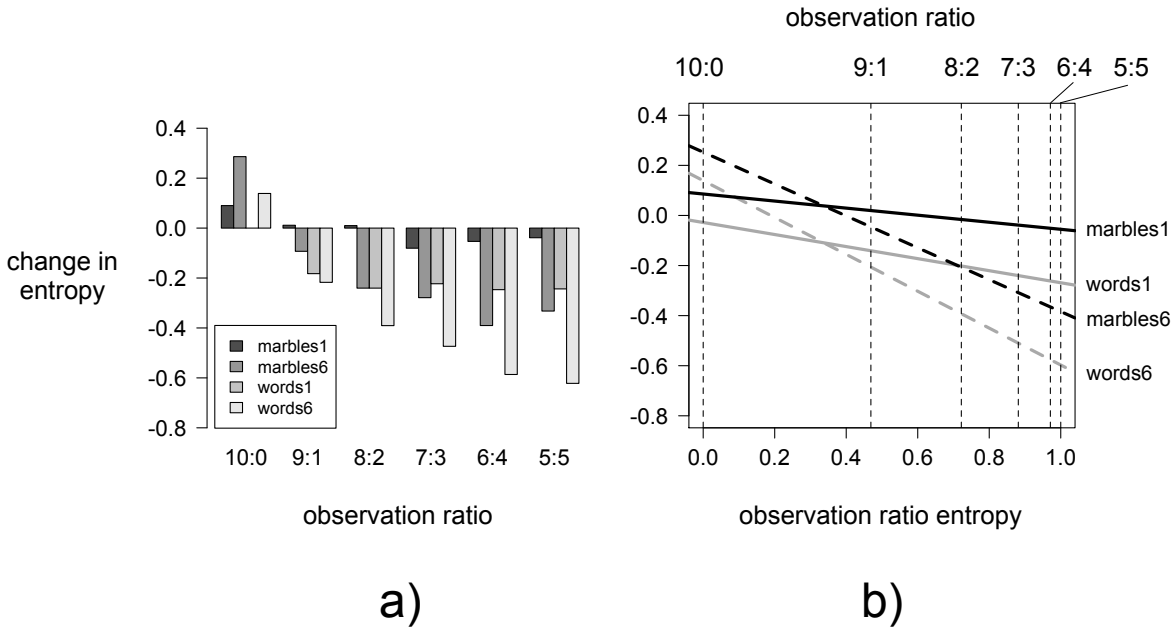


Figure 3.14: a) Average change in entropy per observation ratio, per experimental condition. b) The regression lines for the data plotted to the left according to the best-fit model B.

To understand the directional effects of domain, concurrency, and observation ratio on regularization behavior, Figure 3.14a plots the entropy drop per observation ratio, broken down by the four experimental conditions. Figure 3.14b plots the regression lines for the data in Figure 3.14a according to the best-fit model B. I will explain the directional effects in terms of the regression lines, but Figure 3.14a is provided so that these trends can be crosschecked on the raw data. The regression analysis output in R for model B is provided in Table 3.4, giving the estimates, standard errors, and t values of each parameter in model B. The estimates in this table dictate the intercepts and slopes of the regression lines plotted in Figure 3.14b. The intercept where observation ratio entropy equals 0 for *marbles1* is given by the intercept estimate, 0.08567. This corresponds to the mean change in entropy for participants who observed the 10:0 ratio in *marbles1*. The other intercepts where observation ratio entropy equals 0 are 0.08567 + 0.16715 for *marbles6*, 0.08567 - 0.11373 for *words1*, and 0.08567 + 0.16715 - 0.11373 for *words6*. The slopes of the regression lines are dictated by the estimates on observation ratio and the two interactions with observation ratio. The slope of *marbles1* is -0.14094, *marbles6* is -0.14094 - 0.49568, *words1* is -0.14094 - 0.09968, and *words6* is -0.14094 - 0.49568 - 0.09968.

If there were no effect of observation ratio on entropy change, then all of the regression lines would be horizontal. Here, perfect probability matching behav-

parameter	estimate	standard error	t value
1) intercept (*marbles1*)	0.08567	0.04321	1.983
2) concurrency (multiple)	0.16715	0.04740	3.527
3) domain (words)	-0.11373	0.04149	-2.742
4) observation ratio entropy	-0.14094	0.05494	-2.565
5) interaction between 2 & 4	-0.49568	0.05622	-8.816
6) interaction between 3 & 4	-0.09968	0.04596	-2.169

Table 3.4: Summary of fixed effects for the best-fit model B. The estimates determine the intercepts and slopes for the regression lines plotted in Figure 3.14b.

ior would be a horizontal line because this would lead to no change in entropy despite the observation ratio seen. Therefore, the slope of the line is a measure of divergence from probability matching behavior. If the line is sloped and below zero, then this indicates regularization. If the line is sloped and above zero, then this indicates that participants are producing more variable ratios than the ones they've observed. The significant effect of observation ratio is due to all of these lines being significantly sloped. The result that more regularization occurs for the observation ratios with higher entropy is largely due to the different ceilings on entropy change per observation ratio. Because regularization behavior is necessarily constrained by the level of variation in the input, this is an important topic for all analyses of regularization behavior. In this experimental design, the same uniform distribution of observation ratios is used in each experimental condition and therefore data can be meaningfully compared because they are all subject to the same baseline constraints on entropy change.

If there were no effect of domain, then the regression line for *marbles1* would be on top of *words1*, and the line for *marbles6* would be on top of *words6*. However the words domain lines are lower than those of the marbles domain, meaning that participants regularize more in the words domain than in the marbles domain. This is the significant effect of domain. Likewise, if there were no effect of concurrency, then the lines regression line for *marbles1* would be on top of *marbles6*, and the line for *words1* would be on top of *words6*. However, the 6-item lines are lower (on average) than those of the marbles domain. This means that participants regularize more in the 6-item conditions and constitutes the significant effect of concurrency.

As for the significance of the interactions, there would be no interaction between concurrency and observation ratio if the slope of *marbles1* and *marbles6* were identical, and the slope of *words1* and *words6* were identical. Otherwise, observation ratio would be additively affecting entropy drop by shifting the lines up and down, without changing the slope. However, this is not the case: the

slopes vary within domain. Concurrency and observation ratio jointly modulate entropy change by increasing the regularization of variable observations ratios and increasing the variabilization of regular observation ratios, compared to the one-item conditions. This constitutes the significant interaction between concurrency and observation ratio. Likewise, there would be no interaction between domain and observation ratio if the slope of *marbles1* and *words1* were identical, and the slope of *marbles6* and *words6* were identical. Here, we see that they are nearly parallel. This is the interaction that was weakly supported by the set of best-fit models. And lastly, the null result for an interaction between domain and concurrency is seen in the fact that the two regression lines for the words domain have the same relative slopes as the two regression lines for the marbles domain. This indicates that there is no interaction between domain and concurrency.

Self-reported generating ratio estimates

Although participants' production ratios are a proxy for their estimate of the ratio of the marbles in the containers, or the relative frequency of the synonyms in the language, participants were also asked about these estimates directly.

Figure 3.15 shows the estimated ratios, plotted in the same way that production ratios were plotted in Figure 3.12. Participants estimated the relative percentages of the two variants per item. The x-axis reports the estimated percentage of the majority variant in terms of a probability. The effects of the experimental manipulations on ratio estimates were analyzed in the same way as the production ratio data: in terms of entropy change. The entropy of each observation ratio was subtracted from the entropy of its estimate to yield the change in entropy for each observed and estimated pair of ratios. Figure 3.16 shows the average change in entropy of estimated ratios per condition. The same linear mixed effects regression analysis from the previous section was conducted to obtain significant differences from zero. The estimated ratios of *marbles1* are significantly different than zero, biased toward variability, ($t(1152) = 2.286, p = 0.22$), but none of the other conditions are significantly different from zero: *words1* ($t(1152) = 0.780, p = 0.44$), *marbles6* ($t(1152) = 1.835, p = 0.07$), and *words6* ($t(1152) = -1.458, p = 0.15$). In no condition was a bias toward regularity evident in participants' estimated ratios.

Table 3.5 shows the pairwise comparisons between conditions. There are two significant differences: significantly more regular ratios are estimated in *words6* than in *marbles1* ($t(1152) = -2.679, p = 0.007$) and *marbles6* ($t(1152) = -2.329, p = 0.02$). However, this difference is due to the marble drawing con-

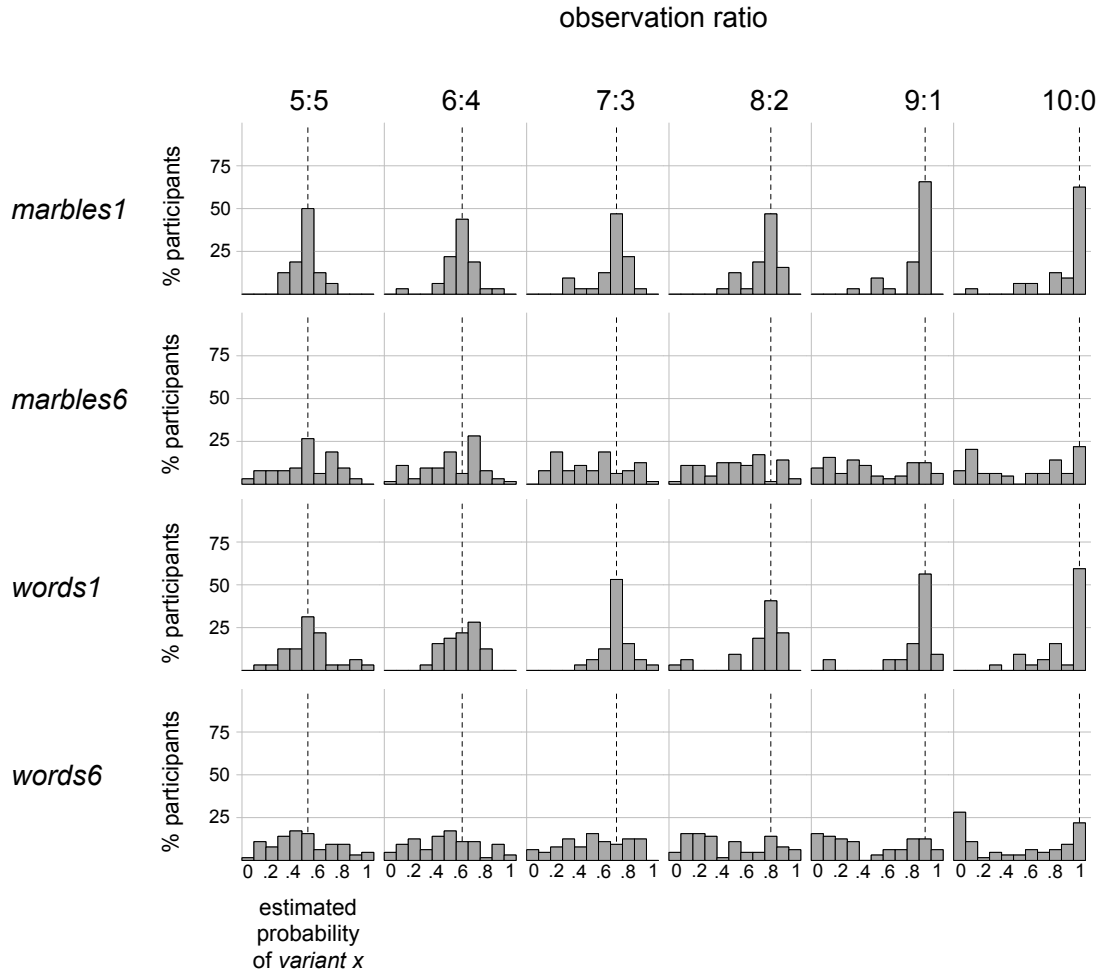


Figure 3.15: Participants estimates of the generating ratio for the four main conditions in Experiment 2. Each row shows the results of one experimental condition. Each column corresponds to one of the six observation ratios, ranging from 5:5 (on the left) to 10:0 (on the right). Each pane contains the distribution of participants estimates per observation ratio. These estimated ratios are displayed on the x-axis as the proportion of *variant x* in their estimate. *Variant x* corresponds to whatever marble/word was in the majority during the observation phase. All observation ratios are indicated by a dashed line. For example, the top left panel gives the results for the 32 participants in condition *marbles1* who observed a 5:5 ratio of marbles. Here, we see that 50% of these participants reported a 5:5 generating ratio, 19% reported a 4:6 generating ratio, and 13% reported a 6:4 ratio.

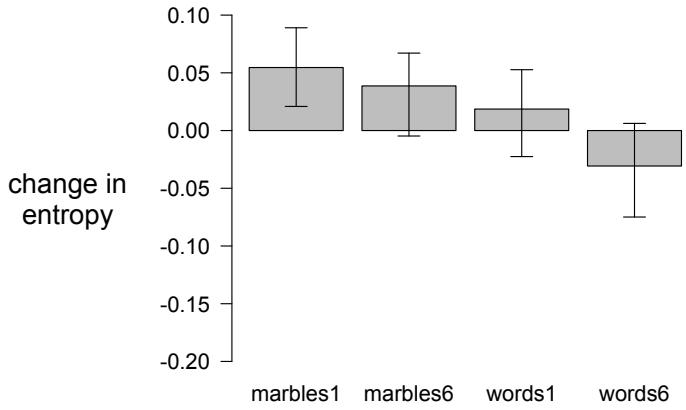


Figure 3.16: Average change in entropy of estimated ratios.

marbles1	*marbles6*	$t(1152) = -0.500, p = 0.62$
	words1	$t(1152) = -1.065, p = 0.29$
	words6	$t(1152) = -2.679, p = 0.007$
marbles6	*words1*	$t(1152) = -0.629, p = 0.53$
	words6	$t(1152) = -2.329, p = 0.02$
words1	*words6*	$t(1152) = -1.549, p = 0.12$

Table 3.5: Significance of pairwise comparisons between the four conditions in Experiment 2.

ditions being more variable, because there was no significant regularization of estimates in *words6*.

The same linear mixed effects regression analysis from the previous section was conducted on the full model reported in Table 3.2, where the dependent variable is change in entropy of the ratio estimates (rather than the production ratios). The same independent variables are significant predictors of entropy change of the estimated ratios. There was a significant effect of domain $\chi^2(4) = 11.735, p = 0.02$, concurrency $\chi^2(4) = 34.916, p < .001$, and observation ratio $\chi^2(4) = 562.04, p < .001$. The direction of this concurrency effect has variable ratios being estimated more often in the one-item tasks than the six-item tasks. And the direction of the domain effect has the marble drawers estimating more variable ratios than the word learners. This means that participants may be encoding frequencies differently due to the concurrency manipulation and due to the domain manipulation. All possible interactive and additive effects from Table 3.3 were assessed as well. An ANOVA showed the same set of models to be the best fits: B ($\chi^2(1) = 27.9293, p < .001$), D ($\chi^2(0) = 27.1800, p < .001$), and F ($\chi^2(0) = 27.2416, p < .001$). Whereas B was the best-fit model of these three

for the production ratio data, none of these three models was significantly better for the estimated ratio data. Participants' production ratios and self-reported estimates of the generating ratio are modulated by the same fixed effects.

In conclusion, there is no evidence that the errors in participants' estimates of the generating ratio are biased toward regularity in any of the conditions where production ratios are regularized (*words1*, *marbles6*, *words6*). This suggests that regularization behavior is not caused by a regularization bias during encoding, but instead caused by something on the recall and production end of this experiment. This explanation seems to fit the *words1* condition best, where participants' estimates accurately reflect the observed ratio, but their productions are regular. In both of the six-item tasks, however, participants' estimates were quite noisy (refer back to Figure 3.12). This indicates that it was more difficult to estimate the generating ratios in six-item task than in the one-item task. Although this difference could very well be due to the higher cognitive load in the six-item task, it is slightly concerning that these estimates seem to carry no information about the observation ratio, whereas participants' productions in the six-item task do. In other words, participants tend regularize by over-producing the majority marble, but seem to have no knowledge about the relative frequencies of the two marbles when it comes to the estimation part of this experiment. This could be due to a greater lack of attention toward the end of the six-item task compared to the shorter, one-item task. Or it could be that the format for eliciting estimates in the six-item condition was more difficult or confusing than the format used for the one-item condition (refer to the instructions and screen shots in Appendix A). If this is the case, then this means that the six-item estimation data is of low quality. However, the regression analysis on estimates did capture the same effects and interactions as were present in the high-quality production data.

Self-reported observation and production frequencies

In the previous section, we saw that participants were worse at estimating ratios for the six-item task than for the one-item task, and this is likely to be due to the six-item task being more difficult, though the exact source of this difficulty is unclear. We also saw that domain is still a significant predictor of entropy drop in estimated ratios. It is possible that differences in frequency estimates are due to domain-specific biases, but this could also be due to a difference in memorability of the marbles versus words stimuli. The two additional questions asked in the one-item tasks shed light on this difference. Participants reported how many times they observed and produced each variant. This is a more concrete question

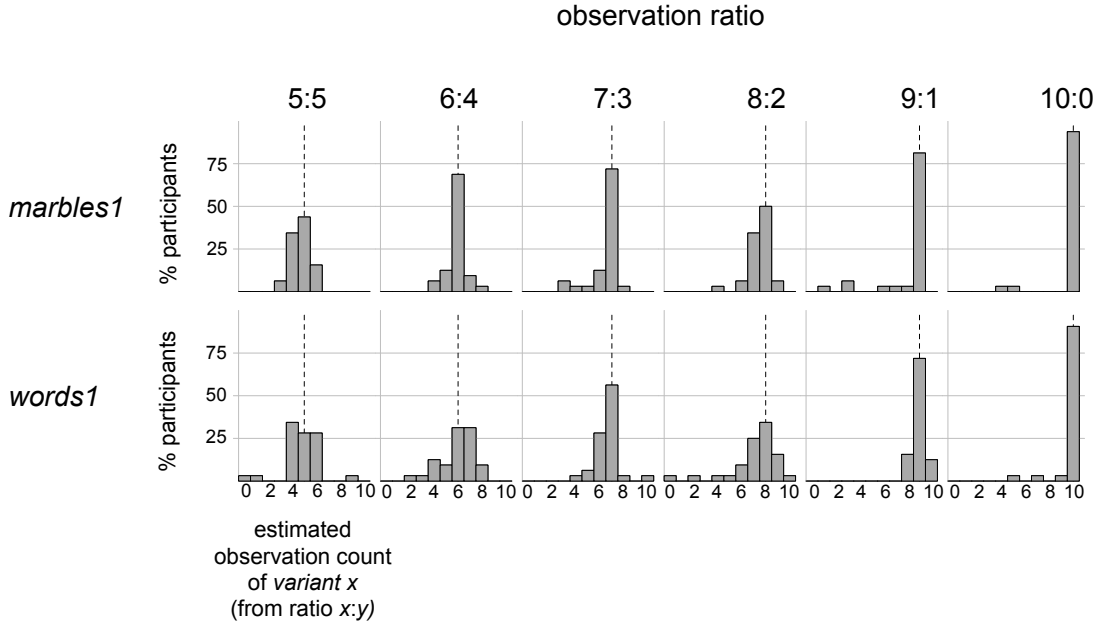


Figure 3.17: Participants' estimate of the number of times they saw *variant* x in the observation phase. The dashed line shows the true number of observations.

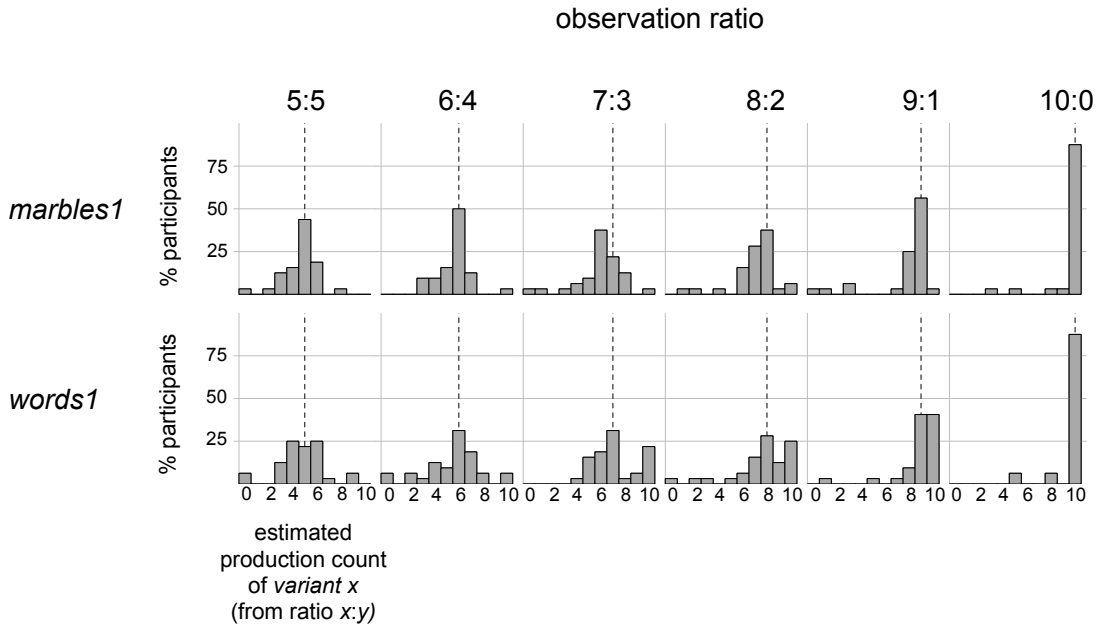


Figure 3.18: Participants' estimate of the number of times they produced *variant* x in the production phase.

that should avoid domain-specific biases that play into participants' estimates and reveal more basic differences in participants' awareness of stimuli frequencies across the two domains.

Figure 3.17 plots the number of times participants reported observing the majority variant, per observation ratio and per condition (*marbles1* or *words1*). Participants in both conditions report their observation ratio with high accuracy. Figure 3.18 plots the number of times participants reported producing the majority variant, per observation ratio and per condition (*marbles1* or *words1*). These data are similar to the distributions produced by participants in each condition. In general, marble drawers recall the probability matching productions they made and word learners recall the regular productions they made.

The question of interest here is whether or not there is a significant difference in recall accuracy across domain. A linear model was used to investigate recall error as a function of domain and observation ratio[12]. For observation recall, a model was constructed where domain and observation ratio were the independent variables, and the dependent variable was the change in entropy between the observed ratio and the ratio each participant recalled observing. For production recall, the dependent variable was the change in entropy between the ratio participants produced and the ratio they recalled producing. Neither model revealed a significant effect of domain. There was no significant difference in recall of observation ratios ($t(380) = -0.392, p = 0.7$) or recall of production ratios ($t(380) = 1.346, p = 0.18$). There was a marginally significant effect of observation ratio on the observed ratio recall ($t(380) = -2.327, p = 0.20$), with participants recalling observations for the more deterministic ratios better. However, there was no significant effect of observation ratio on production ratio recall ($t(380) = -0.591, p = 0.56$), as any effect there would be indirect. This means that basic differences in the attention and recall of marble and word stimuli frequencies are not biasing participants toward different levels of variability or regularity across domains.

3.2.3 Discussion

Both of the experimental manipulations, domain and concurrency, bring participants away from the probability matching behavior of *marbles1* and elicit regularization. Furthermore, there was no evidence that these two sources of regularization behavior interact: they appear to be independent contributors to

[12] A linear mixed effects regression analysis could not be used here because the data points in these conditions are independent observations (i.e. one per participant)

the full amount of linguistic regularization elicited in condition *marbles6*. This independence indicates that different cognitive mechanisms underpin regularization behavior in these two manipulations. However, these mechanisms seem to give rise to fairly similar regularization profiles in which the majority variant is over-regularized. Although Occum's razor would suggest that these two sources of regularization are one and the same, the analysis of interactions revealed strong support for an interaction between the concurrency manipulation and observation ratio, but no strong support for an interaction between the domain manipulation and observation ratio. This means that although word learners do appear to be using frequency information to regularize with the majority variant, their regularization behavior does not seem to be sensitive to the fine-grained difference in stimuli frequencies from different observation ratios. In the non-linguistic domain, participant regularization behavior shows significant differences with respect to the specific observation ratios of the stimuli. This difference suggests that the concurrency manipulation is closely tied to lower-level cognitive processes and memory constraints that operate directly on encoded frequency information, whereas the linguistic regularization bias may be the result of a higher-level decision-making process that uses encoded frequency information in a broader sense. If this is the case, then the behavior of participants in linguistic regularization experiments may be more sensitive to aspects of task framing, including participants' perceived goal of the production phase, than participants in the basic frequency learning experiments of classic psychoeconomics.

Further analysis of participants' self-reported estimates of the generating ratios showed that their regularization bias is not rooted in their estimates. In each of the three conditions where participants regularize their productions (*marbles6*, *words6*, and *words6*), participants' estimates were not significantly biased toward regularity or variability: they were fairly accurate. Although these estimates were much noisier in the six-item tasks, indicating that the six-item task is associated with higher cognitive load, these errors in encoding are not likely drivers of regularization since they themselves are not biased toward regularity. This finding speaks to the results of Perfors (2012), which manipulated cognitive load during linguistic frequency encoding in the observation phase, but found that participants did not regularize. My results have shown that the higher cognitive load of the six-item conditions did not lead to biased encoding errors, and this may also be the case for Perfors (2012). If encoding errors do not drive regularization, then the drivers may lie in memory recall and production. This would also explain the fact that participants in Perfors (2012) did not regularize, but the participants in Experiment 2 did, because the production phase of Experiment 2

also entailed high cognitive load (via concurrent frequency production), whereas the production phase for Perfors (2012) did not. Also, domain was a significant predictor for the estimated ratios data, with word learners yielding less of a variability bias than marble drawers. However, this difference between domain is not due to a difference in participants' attention or ability to recall the marbles versus words stimuli: participants' recall accuracy of the number of times they observed and produced each variant were not different between domains. This is evidence that learners only *estimate* word stimuli frequencies slightly differently than marble stimuli frequencies. But it is not clear that these differences drive regularization behavior, because the estimates showed no regularity bias. Therefore, regularization due to domain may also fall largely on the production side of the task.

In addition to Perfors (2012), two other well-known papers provide experiments in which adult learners do not regularize linguistic input. Hudson Kam and Newport (2005) contrast adult and child learners and conclude that child learners regularize, whereas adult learners do not. However, their assessment of regularization is more coarse-grained than that provided in this chapter, because they group participants either as full regularizers, or variable users. Participants who partially regularize, by making lower-entropy productions than their observations, would fall into the "variable users" category. They find that about 15% of their participants fully regularize. For comparison, in the three conditions of Experiment 2 where participants clearly regularize, the percentage of participants who fully regularize are 29% in *words1*, 33% in *marbles6*, and 59% in *words6*. The linguistic stimuli in Hudson Kam and Newport (2005) is far more complex than that of Experiment 2, and this may be the source of regularization differences here, but in relation to the data in this thesis, a 15% incidence in full regularization behavior is by no means evidence that adult learners do not contribute to regularity in languages, as Hudson Kam and Newport (2005) conclude. Reali and Griffiths (2009) also conclude in one experiment (which *words6* is a replication of) that no regularization bias is evident in the behavior adult learners, whereas when the task is iterated, regularization behavior becomes evident because cultural transmission amplifies the behavioral consequence of participants' weak regularization bias. However, the participants in their first experiment are in fact regularizing, but their population-level analysis of regularization could not detect this (refer back to Figure 3.2 and its accompanying discussion). This point was thoroughly discussed in the literature review at the beginning of this chapter, but it is worthwhile to mention again here that the results of Reali and Griffiths (2009) do not conflict in any way with the results of *words6* in Experiment 2.

In the following sections, Experiments 3 and 4 further explore the issues of domain-general regularization due to concurrent frequency learning and memory constraints, via small modifications to Experiment 2. In the next chapter, Experiment 5 explicitly addresses the production side of this task: it explores the decision-based strategies that may give rise to linguistic regularization when two partners use shared frequency information to achieve coordination in a non-linguistic task.

3.3 Experiment 3: a closer look at domain-general drivers

The previous experiment identified two independent sources of regularization in an artificial language learning task: a domain-general component due to concurrent frequency learning, and a domain-specific component due to the use of linguistic stimuli. Experiment 3 digs deeper into the domain-general aspects of regularization and asks why participants regularize when learning about several frequencies at once.

The artificial language learning experiments reviewed at the beginning of the chapter are all concurrent frequency learning tasks that cover different numbers of choice alternatives of the linguistic variable of interest: for example, Smith and Wonnacott (2010) had two determiners, Hudson Kam and Newport (2009) looked at 0, 4, 6, 10, and 18 determiners, Culbertson et al. (2012) had four word order combinations, and Reali and Griffiths (2009) had 12 lexical items (2 per object). It is likely that both the entropy of these distributions and the number of choice alternatives modulates participants' regularization behavior, but these two factors are difficult to disentangle.

Extensive literature in psychoeconomics provides conflicting results on how the number of choice alternatives modulate probability matching behavior. Gardner (1957, 1958), Cotton and Rechtschaffen (1958), and McCormack (1959) found that presenting more than two choice alternatives leads participants to regularize by overproducing the higher-frequency alternative. However, Detambel (1955), Wittig and Weir (1971), and Weir (1972) find the opposite result: the variability of participant responses increases when learning about more than two choice alternatives. Weir (1972) points out that the studies reporting regularization behavior used a similar feedback schedule in which participants were informed of the correct choice after each testing trial, whether or not they made the correct choice. And in the studies that report probability matching or increased variability with number of choice responses, feedback was only given when participants made the correct choice. Otherwise, they were aware they made an error, but were not told what the correct response was. Because Experiment 2, along with most artificial language learning experiments, gives no feedback and is not an explicit per-trial prediction task, it is not possible to make straightforward predictions from the psychoeconomics literature for the present experiment. However, this literature does show that regularization behavior, of the type observed in artificial language learning tasks, does occur in non-linguistic frequency learning and can

be modulated by things related to cognitive load, like the number of items one must track.

So, why do participants probability match in the one-item marble drawing task, but regularize in the six-item marble drawing task? From the linguistic regularization literature, we know that different levels of regularity in the training data modulate participant regularization behavior. But why do participants regularize at all? Is this a direct response to instances of regular input data, or is it just because they are learning about several items at once?

Experiment 3 addresses these questions by replicating the concurrent frequency learning condition *marbles6* with a set of observation ratios that contain no regularity: {5:5, 5:5, 5:5, 5:5, 5:5, 5:5} This maximally variable set of observation ratios has an average entropy of 1 bit. As far as I am aware, no study has been published in the psychoeconomics literature on probability matching where participants observe all choice alternatives with equal frequencies. In the linguistic regularization literature Culbertson et al. (2012) is the only example, reporting a control condition in which participants observe four word order types with equal frequencies. Here, participants probability matched (mean responses were not significantly different than the input frequency). However in the other conditions, when the input was skewed in a 7:3 ratio, participants did regularize. This suggests that participants may only regularize when there is some evidence of regularity in their input. The authors point out that the control condition data seems to indicate an apparent substantive (i.e. direct) bias toward overproducing three of the four possible orderings in the raw data. So a small regularization bias may be at play, but this did not lead to a significant difference from the input frequency, possibly due to the small sample size of 13 participants in this condition.

If participants in Experiment 3 probability match in response to a maximally variable training set, then this means that the presence of regularity in the observation ratios of Experiment 2 was the driving force behind regularization in the concurrent frequency learning condition *marbles6*. But if participants regularize, this means that there is something about multiple frequency learning itself that elicits regularization behavior. I chose to conduct this new experiment in the non-linguistic domain to see if concurrent frequency learning *without evidence for regularity* can elicit regularization behavior in a domain where participants naturally probability match given two choice alternatives.

Another interesting question that this new observation ratio set raises is about the independence of the individual items being learned. If participants do regularize the maximally variable set, do they regularize it the same amount as

they do the 5:5 ratio from *marbles6*, or do the other ratios in the observation set modulate it somehow? Although languages contain individual, variable units that people can make judgements about, such as a relative frequency of two past tense inflections, languages are learned as systems and sets of frequencies may be learned as systems as well. Learners may make higher-order generalizations about the input they observe, such as "all containers like these are filled with a 50/50 mix of marbles" (e.g. Kemp et al., 2007). In this sense, the maximally variable observation ratio set is regular at a higher level: it contains ratios of all one type. It is possible that multiple ratios of one type reinforce one another, systematically. To supplement this investigation, I chose to test one more observation set: a maximally regular set of all 10:0 ratios. Participant responses to this set will be compared to responses from the 10:0 ratio of *marbles6*.

3.3.1 Method

Participants

68 participants were recruited via Amazon's Mechanical Turk crowdsourcing platform and completed our experiment online. Participant location was restricted to the United States of America and verified by a post-hoc check of participant IP address location. 4 participants were excluded on the basis of the following criteria: failing an Ishihara color vision test[13] (0), self-reporting the use of a pen or pencil during the task[14] (1), not reporting their sex or age (0), or having previously participated in this or any of my experiments, as determined by their user ID with MTurk (2). More participants were recruited than necessary with the expectation that some would be excluded by these criteria. Once the predetermined number of participants per condition was met, the last participants were excluded, totaling 1 participant across all conditions. All excluded participants received the full monetary reward for the task, which was 0.60 USD. The average time taken to complete the experiment was 11 minutes and 17 seconds, with a standard deviation of 2 minutes and 8 seconds. Of the final 64 participants, 50% are female and the mean age is 35.1 (min = 20, max = 69) with a standard deviation of 11.7 years. The breakdown of participants in each of the two conditions (further described in section 3.3.1) is as follows:

[13]I used two plates from the Ishihara color vision test: plate 4 which tests for red-green color deficiency, and plate 23 which tests for protanopia and deuteranopia (see Appendix A.2.1). Participants were excluded if they gave an incorrect answer for one or both of these plates.

[14]In an exit questionnaire.

Conditions:	all	*all 5:5*	*all 10:0*
Participant count	64	32	32
Age: mean	35.1	37.8	32.4
minimum	20	20	20
maximim	69	69	51
standard deviation	11.7	13.3	9.1
Sex (% female)	50%	65%	35%

Materials

The materials, stimuli, and presentation are identical to those in Experiment 2.

Conditions and Design

There are two conditions in this experiment, one in which participants receive only 5:5 observation ratios, which will be called the *all 5:5* condition, and one in which participants receive only 10:0 observation ratios, which will be called the *all 10:0* condition. These two conditions will be compared to the *marbles6* condition from Experiment 2.

condition	set of observation ratios
marbles6	{5:5, 6:4, 7:3, 8:2, 9:1, 10:0}
all 5:5	{5:5, 5:5, 5:5, 5:5, 5:5, 5:5}
all 10:0	{10:0, 10:0,10:0,10:0,10:0,10:0}

Procedure

This is identical to the procedure used in *marbles6* of Experiment 2.

3.3.2 Results

Summary of the ratios produced

Figure 3.19 shows the results of Experiment 3, along with the results of *marbles6* from Experiment 2, reprinted for comparison.[15] Participants in the *all 5:5* condition do not appear to be probability matching on a 5:5 production ratio. Instead,

[15]In this chapter, all of the 6-panel plots like this are organized by observation ratio with randomized container and marble color stimuli within each panel. Because the panels in this experiment cannot be distinguished by observation ratio, I chose to group them by container stimuli (but there was no significant effect of container on regularization behavior). The containers for each plot, from left to right, are the box, basket, jar, bucket, bowl, and pouch.

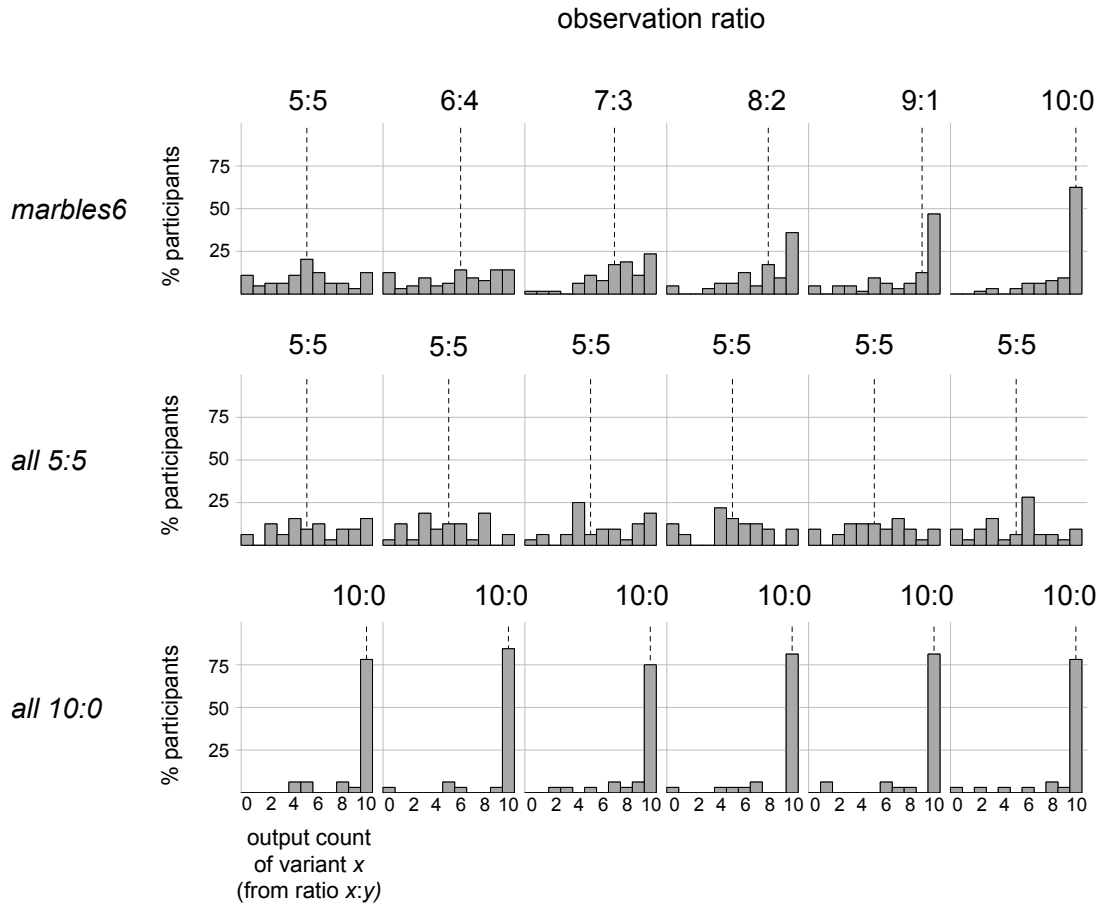


Figure 3.19: Results of the two conditions in Experiment 3. In the first row, *marbles6* is reprinted from Figure 3.12 for comparison. Each row shows the results of one experimental condition and each column corresponds to one of the six observation ratios. Each panel contains the distribution of ratios that participants produced in response to one observation ratio. These production ratios are displayed on the x-axis as the number of times a participant produced *variant x* from the observation ratio *x:y*. *Variant x* corresponds to whatever marble/word was in the majority during the observation phase. All observation ratios are indicated by a dashed line. In *marbles6* there were 64 participants, and thus 64 data points (one from each participant) in each panel. In Experiment 3, there were 32 participants in each condition and thus 32 data points (one from each participant) in each panel. For example, the middle left panel gives the results for 32 of the production ratios in condition *all 5:5*. Here, we see that 9% of these participants also produced a 5:5 ratio, 16% produced a 4:6 ratio, 13% produced a 6:4 ratio.

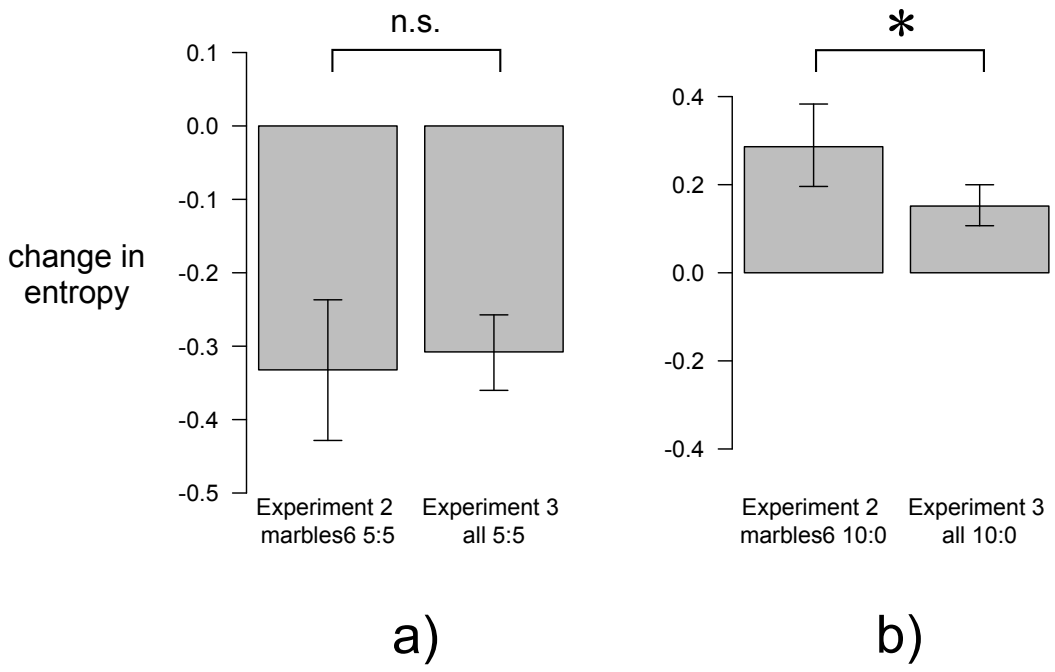


Figure 3.20: Average change in entropy of participants' productions compared to their observation ratio. a) Comparison between the 5:5 ratio from *marbles6* (from Experiment 2) and all of the data from the *all 5:5* condition of Experiment 3. b) Comparison between the 10:0 ratio from *marbles6* (from Experiment 2) and all of the data from the *all 10:0* condition of Experiment 3.

they seem to regularize their responses, often by producing fully regular responses in the 10:0 and 0:10 ratio. In the *all 10:0* condition, participants appear to be maintaining the 10:0 ratio in their productions better than they do for the 10:0 observation ratio in *marbles6*.

Effect of the different sets of observation ratios

Rather than looking at the overall entropy drop per condition, the comparisons of interest here are between the conditions in Experiment 3 and their corresponding ratio in *marbles6*. In the case of the *all 5:5* condition, we want to know if the amount of regularization that occurs when a 5:5 ratio is observed in the context of only other 5:5 ratios is significantly different from when it is learned in the context of a diversity of other ratios, as it is in *marbles6*. Likewise, we want to know if responses to a 10:0 ratio differ between the *all 10:0 marbles6* conditions.

Figure 3.20a shows the average change in entropy for the *all 5:5* data and the data from *marbles6* for the 5:5 observation ratio only. Figure 3.20b shows the average change in entropy for the *all 10:0* data and the data from *marbles6* for

the 10:0 observation ratio only. Because the 5:5 observation ratio is already the maximum entropy of 1 bit, entropy can only stay the same or go down for the two data sets in Figure 3.20a. Likewise for Figure 3.20b, the 10:0 observation ratio is already the minimum entropy of 0 bits, so entropy can only stay the same or go up for these two data sets. Error bars indicate the 95% confidence intervals and were computed with the bootstrap percentile method (Efron, 1979).

To determine whether or not the difference in regularization in Figure 3.20a is significant, I performed a linear mixed effects regression analysis. Entropy change was entered as the dependent variable and condition,*marbles6* (5:5) vs *all 5:5*, was entered as the fixed effect. Participant was entered as a random effect as random intercepts. There was no significant difference between regularization in these conditions ($t(254) = 0.341, p = 0.73$). Participants do not regularize any less or any more when learning about 5:5 ratios in the context of other 5:5 ratios than when they are learning about a 5:5 ratio in the context of many different ratios.

I performed another linear mixed effects regression analysis on the 10:0 observation ratio data as well, to determine whether or not the difference between conditions in Figure 3.20b is significant. Entropy change was entered as the dependent variable and condition, *marbles6* (10:0) vs *all 10:0*) was entered as the fixed effect. Participant was entered as a random effect as random intercepts. There was a significant difference between these conditions ($t(254) = -3.457, p < .001$). Participants maintain more 10:0 ratios when learning in the context of other fully regular ratios than when learning in the context of more variable ratios.

3.3.3 Discussion

Participants regularize just as much when learning about six pairs of marbles in 5:5 proportions as they do when learning about one pair of marbles in a 5:5 proportion that was learned in the context of other, more regular ratios. This means that there is something about concurrent frequency learning itself that causes participants to regularize. Regularization behavior does not necessitate the presence of a majority variant or the presence of some evidence of regularity in the observation set, whether it takes the form of fully regular (10:0) ratios or partially regular ratios with non-zero entropy (ex: 6:4). Additionally, if participants were forming higher-order generalizations about their training set, this did not lead to a significant difference in regularization behavior for the maximally variable observation set.

The maximally regular set, on the other hand, did lead to significantly different

results. When participants were learning about six completely deterministic items (where only one color came out of each container) then participants produced fully regular responses more often than when learning about one 10:0 ratio in the context of other, more variable ratios. This is evidence that the fully regular ratios may be reinforcing each other. However, it may also be evidence that the other ratios in *marbles6* are interfering with the 10:0 item. A higher-order generalization that could interfere with the 10:0 item from *marbles6* may be along the lines of "all these containers seem to be variable, so maybe the one container I only saw one color come out of actually has another color in there, so I should produce some draws of that color." Alternatively, since production trials on the 10:0 item include two possible production choices (one color that was observed and another one that was not), the fact that participants choose the unseen marble more in *marbles6* may be due to an aspect of higher cognitive load in *marbles6* regarding the stimuli numbers. These two data sets may have differed in regularity because participants in the *all 10:0* condition observed 6 marble color stimuli, whereas participants in *marbles6* observed 11 (and 11 unique colors may be harder to remember than 6). There may be a continuum of regularization behavior for experiments using 1 item and 2 variants, up to 6 items and 12 variants, such that stimuli numbers in between the *marbles1* and *marbles6* conditions may elicit interpolated levels of regularization. This point could be addressed in the future by training participants on a three-item task with 6 marble colors, or a one-item task with more than two color variants.

As a general conclusion, this experiment suggests that regularization is triggered by concurrent frequency learning itself and that individual regularization profiles obtained in *marbles6* may be relatively similar to those obtained for any combination of six observation ratios. This has important implications for the analyses in Chapter 6, which address the evolution of frequency distributions from the experiments in this thesis.

If changes in frequencies over time are more determined by the observed ratio of the stimuli than by their specific learning context, then the changes in frequencies of each pair of variants can be modeled independently of one another. However, this experiment also provided evidence that the 10:0 ratio is likely to be a stronger attractor than indicated by the data in *marbles6*. As soon as more than one variant pair enters a 10:0 ratio, they may be more likely to remain in a fully regular state and perhaps, be more likely to convert other items to that state as well. The 5:5 ratio, on the other hand, does not appear to be an attractor that keeps productions in a highly variable state. This implies that highly variable sets of frequencies are less likely to be faithfully transmitted between learners

than more deterministic sets. For now, I will leave these topics to Chapter 6.

This experiment touched on issues of cognitive load and the possibility of memory constraints affecting regularization behavior, but left these questions largely unsatisfied. The following section presents another variation on Experiment 2 that addresses these issues more directly.

3.4 Experiment 4: a closer look at memory

Some of the results we have seen so far suggest that aspects of the production phase may be driving regularization behavior. When participant responses were compared across domain, participants in the one-item conditions were fully aware of the frequencies of marbles/words they observed. However, when it came to the production phase, word learners produced many fully-regular responses whereas marble drawers did not. The difference in participant behavior across these two conditions is likely due to a difference in participants' interpretation of the task. The heavily bimodal distribution of participant responses in *words1*, in which some participants probability matched perfectly and others fully-regularized, corroborates this view. This suggests that some participants explicitly decided to match the frequencies, whereas others may have thought they should choose the "best" word. This relates to task framing and will be explored further in Experiment 5.

We also saw some evidence that participants' biases on frequency estimates were *not* candidate sources of the behavioral bias toward producing regular ratios. As discussed in the literature review at the beginning of this chapter, Perfors (2012) has shown that increasing cognitive load in the observation phase (i.e. during memory encoding) does not lead to regularization behavior. Perfors (2012) also reports Bayesian modeling evidence showing that learners will not regularize under memory constraints unless they possess an inductive bias toward regular ratios in their frequency estimates. This result is corroborated by the frequency estimate data in Experiment 2. Participants under cognitive load in the six-item conditions were worse at estimating the generating ratios behind the data they observed, however this estimate itself was not biased toward regularity. The fact that estimates were much worse in the six-item conditions does show that cognitive load was higher when participants learned about six items concurrently, as opposed to one item on its own. This means that cognitive load may be equally high during the production phase, where participants must produce marbles draws/naming events for several different items concurrently. So perhaps it is the cognitive load involved in concurrent frequency production that drives regularization behavior.

Experiment 4 consists of two manipulations to *words6* in Experiment 2 that are designed to alleviate cognitive load during the observation and production phase. The hypothesis is that alleviating cognitive load in the production phase

only will lead to a significant decrease in regularization behavior, whereas alleviating cognitive load in the observation phase only will not.

3.4.1 Method

Participants

154 participants were recruited via Amazon's Mechanical Turk crowdsourcing platform and completed our experiment online. Participant location was restricted to the United States of America and verified by a post-hoc check of participant IP address location. 26 participants were excluded on the basis of the following criteria: self-reporting the use of a pen or pencil during the task[16] (5), not reporting their sex or age (3), or having previously participated in this or any of my experiments, as determined by their user ID with MTurk (18). All excluded participants received the full monetary reward for the task, which was 0.60 USD. The average time taken to complete the experiment was 10 minutes and 14 seconds, with a standard deviation of 1 minute and 49 seconds. A log was kept of the number of times the experiment was accessed. 67% of all people who accessed the experiment completed it.[17] Of the final 128 participants, 40% are female and the mean age is 34.6 (min = 19, max = 66) with a standard deviation of 11.8 years. The breakdown of participants in each of the two conditions (further described in section 3.3.1) are as follows:

Conditions:	all	*observation in blocks*	*production in blocks*
Participant count	128	64	64
Age: mean	34.6	34.2	35.0
minimum	19	19	19
maximim	66	66	65
standard deviation	11.8	11.6	12.0
Sex (% female)	36%	41%	31%
Completion rate	67%	66%	68%

[16] In an exit questionnaire.

[17] I only began collecting this type of data for Experiment 4, which is the last experiment I ran. At least for this experiment, we can say that participants were not selectively dropping out between conditions because the split was near-even between conditions (66% and 68%, see table above). All of the MTurk experiments reported in this thesis prevented the same IP address from accessing the experiment more than once. Therefore, we can assume that the completion rate refers to unique users and not several attempts from the same user.

Materials

The materials, stimuli, and presentation are identical to those in Experiment 2.

Conditions and Design

There are two conditions which constitute manipulations to the observation and production phases. The details of these differences are presented in the Procedure section below. Both of the conditions in Experiment 4 will be compared to the data from the *words6* condition in Experiment 2.

condition	observation trials	production trials
words6	interleaved	interleaved
observation in blocks	in blocks	interleaved
production in blocks	interleaved	in blocks

Procedure

The procedure used in the *words6* condition of Experiment 2 was modified by organizing trials into blocks per object. This means that each of the 10 training trials per object were presented consecutively. In the *observation in blocks* condition, only the observation phase was in blocks, whereas the production phase remained interleaved and identical to that of *words6* (refer back to the schemas in Figure 3.11). In this condition, for example, a participant would see 10 naming events for one of the objects, then 10 naming events for another one of the objects, and so on until 10 naming events for each of the six objects had been seen. The order in which objects appeared was randomized across participants. In the *production in blocks* condition, only the production phase was in blocks, whereas the observation phase remained interleaved and identical to that of *words6*. In this condition, participants would generate 10 naming events for one item, then 10 more for another item, and so on. The order in which objects appeared was randomized across participants.

3.4.2 Results

Summary of the ratios produced

Figure 3.21 shows the results of Experiment 4, along with the results of *words6* from Experiment 2, reprinted for comparison. The results of both conditions are similar to that of *words6*: participants regularize their responses, they tend

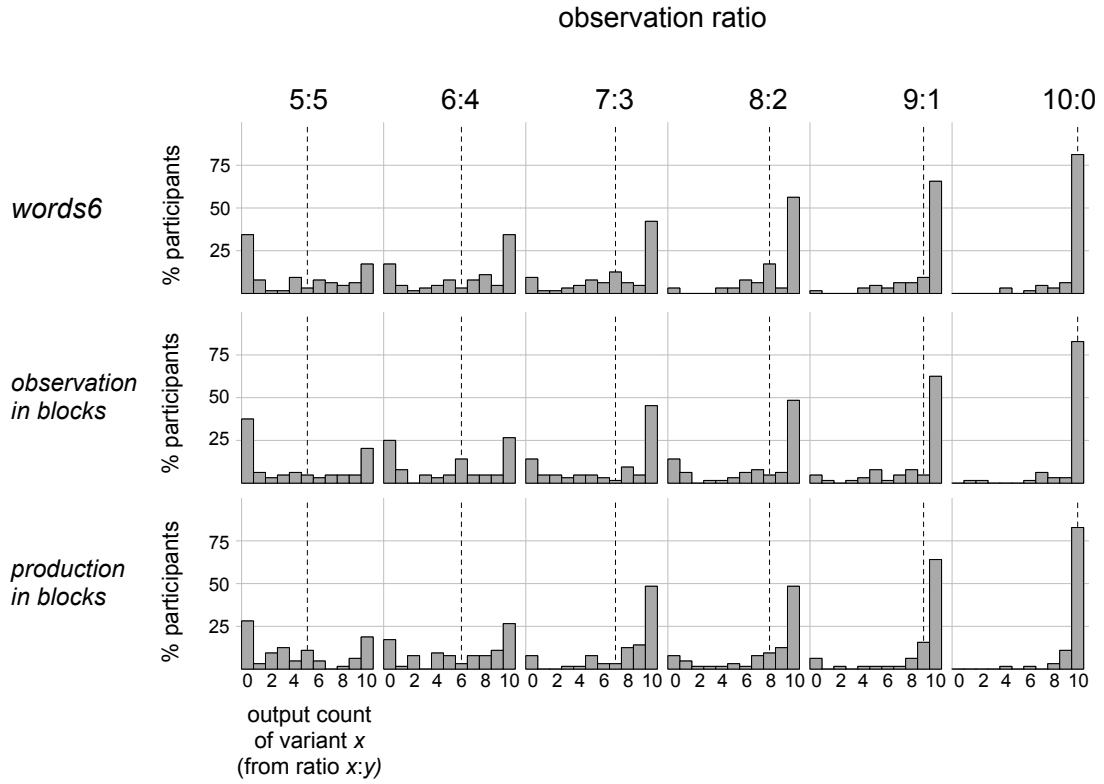


Figure 3.21: Results of the two conditions in Experiment 4. In the first row, *words6* is reprinted for comparison. Each row shows the results of one experimental condition. Each column corresponds to one of the observation ratios, ranging from 5:5 (on the left) to 10:0 (on the right). These production ratios are displayed on the x-axis as the number of times a participant produced *variant x* from the observation ratio *x:y*. *Variant x* corresponds to whatever marble/-word was in the majority during the observation phase. All observation ratios are indicated by a dashed line. For example, the bottom left-hand pane shows the results for the 5:5 observation ratio in the *production in blocks* condition. Here we see that 28% of the participants produced *variant x* 0 times, 11% produced it 5 times (the same frequency in which they observed it), and 19% produced it 10 times. In the 5:5 observation ratio there is no majority variant, but in all of the other observation ratios *variant x* corresponds to the majority variant. In the next pane to the right, *variant x* was observed 6 times. Here, it was also produced 6 times by only 3% of participants, whereas 27% produced it all of the time and 17% did not produce it at all.

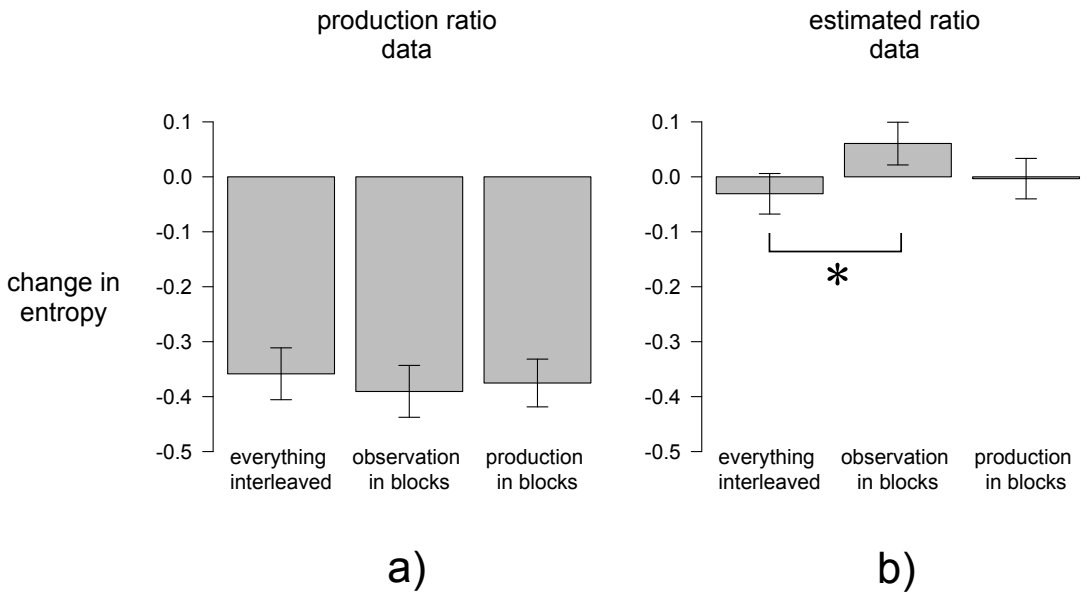


Figure 3.22: a) The average change in entropy of each observation and production ratio pair, per condition. b) The average change in entropy of estimated ratios, compared to observation ratios. Error bars are 95% confidence intervals.

to do so by overproducing the majority word, and this effect is stronger as the frequency of the majority word increases.

Effect of condition on regularization behavior

Figure 3.22a shows the average change in entropy of each observation and production ratio pair, per condition. If the average change in entropy is significantly below zero, then participants are regularizing. Significant differences from zero were assessed by a linear mixed effects regression analysis. A model was constructed with entropy change as the dependent variable and condition as the independent variable (i.e. fixed effect). Participant was included as a random effect as random intercepts. No obvious deviations from normality or homoscedasticity were apparent from a visual inspection of residual plots.

The model was releveled and run two times to obtain the intercept value for each condition. Here, the intercept corresponds exactly to the mean entropy change of the condition and the regression analysis provides a t-statistic to evaluate whether or not this mean is significantly different from zero. Both conditions elicited a significant amount of regularization behavior: *observation in blocks* ($t(766) = -11.180, p < .001$), *production in blocks* ($t(766) = -10.739, p < .001$).

As for differences between conditions, a linear mixed effects regression analysis was conducted to determine whether or not the type of observation or production

regime significantly contributed to the prediction of participant regularization behavior. Entropy change was entered as the dependent variable. Observation regime (interleaved or in blocks), production regime (interleaved or in blocks), and the entropy of the observation ratio were entered as the independent variables (i.e. fixed effects). Participant was included as a random effect (as random intercepts). P-values were obtained by likelihood ratio tests, performed by an ANOVA, on the full model with the effect in question against a reduced model that omits the effect in question. If the full model is significantly better at describing the data than the reduced model, that means the effect in question is a significant predictor of entropy change. The χ^2 test statistic and p-value of the full to reduced model comparison is reported for the effect in question.

There was not a significant effect of blocking production trials on regularization behavior ($\chi^2(1) = 1.0634, p = 0.30$): the full model which included production regime type as a predictor did no better than the reduced model that omitted production regime type. Therefore, alleviating cognitive load during the production phase does not lead participants to regularize significantly less (as hypothesized) or more.

Also, blocking observation trials had no significant effect on regularization behavior ($\chi^2(1) = 0.3817, p = 0.54$): the full model which included observation regime as a predictor did no better than the reduced model that omitted it. If alleviating cognitive load during the observation phase caused participants to encode frequencies differently, this did not lead to a difference in regularization behavior.

Effect of condition on self-reported frequency estimates

Although the blocking of observation trials did not lead to a significant difference in regularization behavior, it still may affect how participants encode the frequencies. Are people regularizing the same because there is no difference in encoding between the two observation regimes? Or are people regularizing for other reasons, which do not heavily involve frequency knowledge?

Participants' self-reported estimates of the generating ratios were analyzed in the same way that they were in Experiment 2. Figure 3.22b shows the average change in entropy between each observation ratio and the participant's estimate of the generating ratio behind that observation ratio, per condition. The results of these data for *words6* are re-printed for comparison (left-most bar). Differences between conditions were assessed by the analysis carried out for Figure 3.22a, but with the relevant dependent variable: entropy change between observation and

estimated ratio.

There was not a significant effect of blocking production trials on participants' estimate of the generating ratios ($\chi^2(1) = 0, p = 1$): the full model which included production regime as a predictor did no better than the reduced model that omitted production regime. This would be expected since the production regime should not effect participants' encoding of frequencies. However, blocking observation trials did have a significant effect on participants' estimate of the generating ratios ($\chi^2(1) = 6.2725, p = 0.01$): the full model which included observation regime as a predictor explained more of the data than the reduced model that omitted it. Participants in the *observation in blocks* condition reported more variation 1) than was actually present in their observation set and 2) than participants in the two other conditions with interleaved observation phases did. This indicates that participants notice more variation in their observation ratios when observation trials are in blocks than when they are interleaved.

3.4.3 Discussion

All together, these results show that alleviating cognitive load during observation allows participants to encode more variation. However this does not affect the extent to which they regularize their responses. Additionally, alleviating cognitive load during production does not lead to a change in regularization behavior, as hypothesized. This may be because the manipulations in this experiment did not change the cognitive load very much. However, these results show that the particular level of regularity that participants produce when learning about six items concurrently[18] is quite robust to these experimental manipulations. This indicates that an important driver in regularization behavior in this task may have to do with participants' perceived goal in this task. As discussed earlier in the analyses of the domain-specific regularization behavior obtained in Experiment 2, *words1*, the linguistic framing of the task may prompt participants to choose the"best" word in their production trials. Therefore, the linguistic framing of this task may have swamped out any nuanced difference in regularization behavior due to the difference in encoded frequencies from the *observation in blocks* manipulation. Additionally, if participants were trying to choose the "best" word in the production trials, this would have precluded them from probability matching their observation frequencies even if it would have been easier for them to do so when production trials were in blocks. On the basis of this, I hypothesize that

[18]Although the trials were not interleaved in the *observation in blocks* condition, participants still learned about all six items before they began the production phase. This still required participants to keep track of frequency information for all of the items at the same time.

if Experiment 4 were run again in the non-linguistic domain with marbles and container stimuli, the *production in blocks* condition may elicit significantly less regularization, because linguistic-domain regularization biases will not be present to override this behavior.

For now, we will leave the investigation of regularization due to memory constraints and turn our attention to the pragmatic factors that may be eliciting regularization behavior in linguistic frequency learning tasks.

Chapter 4

Regularization during non-linguistic coordination

4.1 Introduction

In the previous chapter, we saw how reframing our basic frequency learning task as a word learning task can take participants away from probability matching behavior and elicit regularization. We also saw that this type of domain-specific regularization is more likely to be due to production aspects of this task than to a distortion of the observed frequencies encoded in memory. This is because the linguistic framing of the task prompts participants to *do* something different with the frequencies they remember when they come to the production phase of the experiment. If linguistic-domain frequency learning tasks are easily modulated by task framing, then this may be an important source of the discrepancies found in the linguistic regularization literature (as discussed in Chapter 3).

This chapter presents Experiment 5, in which the production phase of our basic frequency learning task is manipulated to investigate another potentially important source of regularity in language: the pressure to coordinate. Language can be understood as a coordination game (Vanderschraaf, 2014), where the solution is arbitrary (Saussure, 1966), because it does not matter which signal two people use for a meaning, so long as it allows the players to arrive at the same meaning (Lewis, 1969). If language is a coordination game, what coordination strategies do people use in language learning and production? And do players modulate their coordination strategies on the basis of their shared experience with a linguistic system?

Because I am interested in the roots of linguistic regularization behavior, I am going to focus on frequency information as a form of shared experience that

two learners can have about their language. In Experiment 2, we saw that language learners do use their experience of the relative frequencies of lexical items to inform productions, and most often they do so by over-producing the majority variant via regularization. If linguistic regularization behavior has any basis in domain-general coordination strategies, we may expect to see this behavior paralleled in a non-linguistic coordination task. The goal of this chapter is to determine whether two coordinating individuals spontaneously use their shared experience with relative frequencies of non-linguistic stimuli to succeed in a coordination task.

In Experiment 1, I showed that participants probability match in a basic frequency reproduction task and are able to make predictions about the ratio of future events. In the linguistic condition of Experiment 2, I re-framed the task so that participants were making the same kind of predictions, not about marbles in containers, but about words for objects. Although participants learned about the frequencies equally well across domains, many participants in the word production task seemed to fully regularize: they only produced the most frequent word on all production trials. Full regularization with the *observed majority* variant is behaviorally equivalent to maximization: choosing the most frequent outcome on all trials. As discussed in Section 2.1, maximizing is the rational strategy to play in prediction tasks if outcomes are truly random because choosing the most frequent outcome on all trials maximizes the expected payoff. Although the psychoeconomics literature has shown it to be notoriously difficult to get participants to maximize in frequency-based prediction tasks, even with extensive training (Shanks et al., 2002), participants readily maximize in a linguistic frequency learning task. One of the crucial differences in framing between these two tasks is that one involves making predictions about asocial events in your environment, and the other involves making predictions about the behavior of another individual, namely whoever was producing the names for the objects in the observation phase of the word learning task. Only one more step is needed to turn this task into a coordination game: have participants predict the behavior of someone else who is predicting their behavior. The questions here are: Does a coordination game that involves frequency learning also elicit maximization behavior? How is this modulated by the particular frequencies observed? And how does this compare to regularization behavior in the linguistic frequency learning tasks?

To investigate this, we[1] took Experiment 1 and changed the production task.

[1]This experiment was jointly developed and run in the lab with Caroline Kamps for her master project, co-supervised by Simon Kirby and myself. The design of Experiment 5 was the

In the new experiment, Experiment 5, participants observe 10 marble draws from a bag (one at a time, just as in Experiment 1), but instead of being asked to produce several more likely draws to come from this bag, they are asked to coordinate with a partner, who also saw 10 draws from the same bag. In this new production phase, participants are asked to write down a marble color and told that the only goal is to write the same color that their partner writes on their piece of paper. This is repeated 10 times, without feedback, and constitutes a series of one-shot games played between a pair of participants.[2] We will see that this production regime elicits a different profile of responses than those elicited in Experiments 1 and 2, but one that is still dominated by regularization behavior. But before I describe the experiment in detail, I will present the game theory behind this new production task and attempt to link rational behavior in coordination games to regularization behavior in frequency learning tasks.

4.1.1 A game-theoretic analysis of this coordination game

In game theoretic terminology, the production task in Experiment 5 is a tacit coordination game (Schelling, 1960, 54-58). In tacit coordination games, individuals must coordinate their behavior to arrive at the same solution in a situation where communication is incomplete or impossible. This is a form of a pure coordination game, where there is no conflict of interest between players; players want to coordinate and they both receive an equal reward if they succeed. Additionally, the solution in coordination games is arbitrary; any action can be the correct solution, so long as both of the players choose that action. These are either played as one-shot games, or repeated games where feedback may or may not be given.

A similar tacit coordination game to Experiment 5 is called Heads and Tails (Schelling, 1960, 56; Mehta et al., 1994a, 164). Here, two players have to choose between *heads* or *tails* and they know that if they choose the same thing as their partner, they get a reward. Figure 4.1 gives the standard notation of this game as a payoff matrix. If participants choose (*heads*, *heads*) or (*tails*, *tails*), they will each get a payoff of 1. If participants choose (*heads*, *tails*) or (*tails*, *heads*) they will get nothing. Many games, this one included, have one or more

product of many conversations with various members of the Language Evolution and Computation Research Group at the University of Edinburgh, namely Kenny Smith, Nikolaus Ritt, Bill Thompson, Kevin Stadler, Catriona Silvey, James Winters, Matthew Spike, Mark Atkinson, and Justin Quillinan.

[2]It is important to keep in mind that the production phase is *without* feedback, because informing participants of success or failure to coordinate per production trial, or allowing them to see their own past productions, would drastically alter the structure of this game.

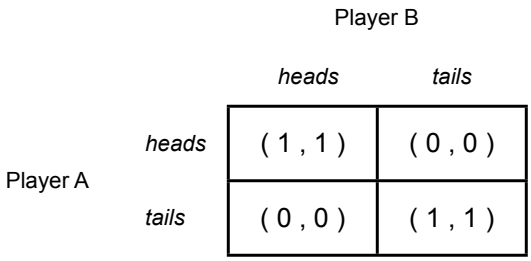


Figure 4.1: A standard example of a pure coordination game.

equilibrium points, which is a set of strategies that rational players will converge upon. In Heads and Tails there are two strategies that a player can take (*heads* or *tails*) and four sets of strategies that the pair of players can take: (*heads*, *heads*), (*heads*, *tails*), (*tails*, *heads*), (*tails*, *tails*). Nash (1950) formalized the concept of an equilibrium for game theory: a particular strategy set is in a Nash equilibrium if neither player can improve their own payoff by unilaterally changing their strategy. In Heads and Tails, the strategy set (*heads*, *tails*) is not a Nash equilibrium because player A can increase their payoff by switching to the tails strategy. Likewise, player B could increase their payoff by switching to the heads strategy. If at least one player can do better by switching to another strategy, that strategy set is not a Nash equilibrium. The strategy set (*heads*, *heads*), however, is a Nash equilibrium, because if either player switched to tails, their payoff would go down.

In Heads and Tails, there are two pure strategy Nash equilibria of (*heads*, *heads*) and (*tails*, *tails*) (Osborne and Rubinstein, 1994). When a player chooses one strategy all of the time, then this is known as a pure strategy. However, a player may also adopt a mixed strategy, which is a probability distribution over pure strategies. Nash (1950) proved that all games with a finite number of players have at least one mixed strategy equilibrium. In Heads and Tails there is one mixed-strategy Nash equilibrium, where heads or tails is played randomly, with equal probability. The expected payoffs for the pure strategies are (1,1) and the expected payoff for the mixed strategy is (0.5,0.5). The probability that two rational players will achieve coordination is this game is 0.5 (Osborne and Rubinstein, 1994). In standard game theory, the payoff values in the matrix constitute all of the information that rational players need to know to make their decisions.

The coordination game of Experiment 5 is equivalent to that of Heads and Tails (Figure 4.2), with an unspecified but positive, equal payoff for each player (further described in Section 4.2.1). In the case of Experiment 3, there are three

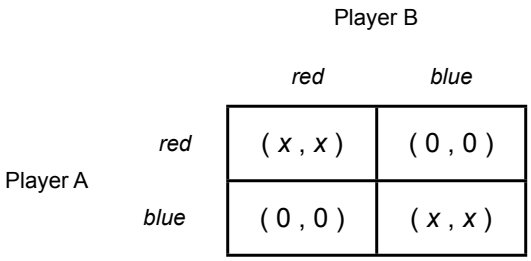


Figure 4.2: The coordination game used in the production phase of Experiment 3, where payoff $x > 0$.

Nash equilibria: two pure strategies (*red*, *red*) and (*blue*, *blue*) and one mixed strategy with a (0.5, 0.5) probability distribution over each pure strategy. If participants in Experiment 5 act rationally, and only use the payoff information to inform their decisions, we would expect them to either choose the red marble in all of their production trials, choose the blue marble in all of their production trials, or play the mixed strategy and produce a 5:5 (or near-5:5) ratio of reds and blues. Additionally, rational participants would be expected to succeed in coordinating no better than chance. However, people use more than just the payoff information to coordinate: they make decisions on the basis of asymmetrical saliencies between their options, known as focal points.

4.1.2 Focal points in coordination games

It is well-known that people are more successful at coordinating in one-shot coordination games than classical game theory can explain (Bardsley and Mehta, 2010). Schelling (1960) carried out a series of informal coordination games and reported that participants regularly coordinated at rates much higher than chance. In the case of the Heads and Tails game, he reported that 86% of participants chose *heads* and this result was replicated in a controlled experiment by Mehta et al. (1994a), where 87% of participants chose *heads*. Schelling proposed the concept of *focal points*, now also known as *Schelling points*, to explain people's uncanny success at solving coordination games. He describes focal points as a class of solutions that have a particular prominence or conspicuousness that "may depend on analogy, precedent, accidental arrangement, symmetry, aesthetic or geometric configuration, casuistic reasoning, and who the parties are and what they know about each other (p.57)." Critically, focal points deal with information exogenous to the payoff structure of the game and thus affect human decision making in coordination games in a way that standard game theory does not account for.

One of the main sources of exogenous information is the *labelling* of the choices (Mehta et al., 1994a; Sugden, 1995), such as *heads* or *tails*, *red* or *blue*, etc. In Heads and Tails, participants seem to share a pre-existing convention or shared cultural knowledge that makes *heads* the preferred choice. Color can also be used as a focal point for successful coordination. Mehta et al. (1994b) show that two colors, red and blue, are strongly preferred as focal points among their participants during a coordination game. In another coordination experiment by Wilson and Rhodes (1997), which required a set of color stimuli that participants were unable to exploit for coordination, the researchers excluded red and blue because "pre-tests indicated that subjects automatically coordinated around these two colors". Scott-Phillips et al. (2009) also show that participants successfully exploit color while trying to solve a coordination game and that red was exploited most often (from a set of four possible colors: red, blue, green, and yellow).

In Experiment 5, all participants see draws of red and blue marbles in the observation phase and may infer that the other participant also saw red and blue marbles. This inference may limit participants' productions to red and blue, although they are free to write any color in the production phase in attempt to coordinate. Because these two colors are the two most common focal colors, their relative difference in saliency may not be large enough for participants to exploit one of these colors as a focal point for successful coordination. I will assess participants' ability to coordinate on the basis of color, but my main interest is in their ability to coordinate on the basis of frequency.

4.1.3 Frequency-based focal points

In addition to color, the shared observation phase introduces another form of labelling that may be exploited for coordination: the *observed majority* and *minority* marble. Several of the games reported in Schelling (1960) and Mehta et al. (1994a) were solved with recourse to *cultural prominence*, which is correlated to frequency. For example, in the game Name a Mountain, most players chose Everest, which is explained (in a somewhat post hoc fashion) as being the most culturally prominent option.

Sugden (1995) formalizes the link between cultural prominence and rational choice via players' use of the relative frequency of options in the environment. In his hypothetical game Name a Word, players must say any word and coordinate on their choice of the word. Assuming that each player has an independent random sample of words (from a particular newspaper they occasionally read), he shows that the decision rule "choose the most frequent item in my sample"

maximizes the expected payoff. The more skewed the veridical distribution is, the more likely players are to successfully coordinate on whatever the most frequent word in their sample was. Many events in the world follow highly skewed distributions, the most ubiquitous example being the power law (Clauset et al., 2009; Mitzenmacher, 2004; Bookstein, 1990), and word frequencies reliably conform to power law distributions (Zipf, 1932). Therefore, the decision rule "choose the most frequent item in my sample" is well-suited to coordinating on words or natural events in the environment.

Experiment 5 serves as an empirical test of this frequency-based decision rule in a non-linguistic domain and I hypothesize that participants will coordinate on the higher-frequency item. To disentangle any effects of coordination on the basis of focal color, I ran this experiment in two conditions: a 5:5 observation ratio and a 7:3 observation ratio. In the 5:5 observation ratio, there is no frequency-based focal point and we can determine whether or not participants successfully coordinate on the basis of color. In the 7:3 observation ratio, there is a frequency-based focal point and this will be counterbalanced so that half of the participants see blue as the majority and half see red as the majority. From this counterbalance, we can see if successful coordination was due more to frequency or to color.

4.1.4 Nash equilibria and regularization behavior

As stated earlier, there are three Nash equilibria strategies in this game: two types of *pure strategy*, where participants either produce all red or all blue, and a *mixed strategy*, where participants choose between red and blue randomly, with equal probability (Osborne and Rubinstein, 1994). Each of these strategies is associated with a production ratio. Participants playing the pure strategy will produce a 0:10 or 10:0 ratio. Participants playing the mixed strategy will be most likely to produce a 5:5 ratio, but with binomial variance due to the random sampling nature of this strategy. Playing either of the pure strategies after observing a variable ratio will lead to regularization behavior as defined in Section 3.1. If participants distinguish between the two pure strategies on the basis of focal-frequency and choose the majority variant on all of their production trials, then this is equivalent to regularizing by overproducing the majority variant (as obtained in Experiment 2) and is also equivalent to maximizing during frequency prediction. Furthermore, if different observation ratios modulate participants' likelihood of playing the pure strategy, then this may result in different levels of regularization per observation ratio condition, as obtained in Experiment 2.

4.1.5 Research questions

The most basic research question here is, can participants coordinate in this task? And if so, do they exploit the marbles' colors and/or relative frequencies as focal points for successful coordination? Of particular interest is how participants exploit frequency. Do participants regularize during tacit coordination? And how does this regularization profile compare to regularization during linguistic frequency learning and production? Since we saw how Nash equilibrium strategies overlap with regularization behavior, we also want to know if participants are achieving the Nash equilibrium strategies and whether strategy choice is modulated by the different observation ratios.

4.2 Experiment 5: regularization biases in coordination

4.2.1 Method

Participants

60 participants were recruited from the University of Edinburgh student community via the university's online participant recruitment system. 37 of the participants were female and the mean age was 23.9 years (min = 22, max = 33) with a standard deviation of 3.2 years. Because the experiment was carried out in pairs, participants either brought a friend along as the second participant, or were paired with another solo respondent. All participants received £4 (the advertized £3 plus a £1 bonus) in monetary compensation for their time. Only participants who self-reported normal color vision were recruited and this was verified by an Ishihara color vision test.[3]

Materials & Stimuli

60 sequences of marble draws were randomly generated in compliance with two observation ratios. The 30 sequences in the 5:5 condition each consisted of 5 blue marbles and 5 red marbles in random order. The 30 sequences in the 7:3 condition were counterbalanced so that 15 sequences were comprised of 3 reds and 7 blues, and 15 were comprised of 3 blues and 7 reds, also in random order. To ensure a

[3]This was the same test used in Experiment 2 and 3 (see Appendix A.2.1). Printouts of the two plates were held up and participants wrote down the number they saw. All participants answered correctly for both plates. One plate tested for red-green color deficiency and the other tested for protanopia and deuteranopia.

representative sample of random sequences, 70% of the sequences began with the majority marble and 70% of the sequences ended with the majority marble. This was done because of primacy and recency effects in learning. Because participants are better at recalling the first and last elements of a sequence (e.g. Deese and Kaufman, 1957), if all of the first and last observation trials happened to be the majority marble, then the data sample itself may bias participants away from a 7:3 representation of their observations. Likewise, the sequences in the 5:5 condition were also made to be representatively random in their first and last elements. 7 strings began and ended with blue, 8 strings began with blue and ended with red, 7 strings began with red and ended with blue, and 8 strings began and ended with red.

These sequences were delivered to participants via a black magician's bag with two internal compartments and a handle, which the experimenter held. One compartment contained approximately 45 red marbles and the other compartment contained approximately 45 blue marbles. The compartments were divided by a piece of black fabric attached to a semi-circular piece of metal that could be flipped from one side of the bag's lip to the other, via a small switch under the handle. The contents of only one compartment can be accessed when someone reaches into the bag.[4] This allowed for the accurate presentation of a controlled set of marble draws across participants. If participants had been making truly random draws from a bag of mixed red and blue marbles, there is no guarantee that the 30 slots in each observation condition would be filled in a finite amount of time. So this departure from the "no misleading participants" culture of experimental economics was done for practical considerations[5]. Participants were fully debriefed after the experiment ended.

Procedure

The experiment consisted of an observation phase and a production phase. Two participants entered the room and sat side by side to fill out their consent forms, take a color blindness test, and receive instructions (see Appendix A.3 for the exact verbal and written instructions). Participants then sat in chairs back to back for the remainder of the experiment. They were instructed to remain silent throughout the experiment and not to engage in any form of communication with their partner. In the observation phase, participants drew marbles from the same

[4]Participants were asked if they noticed anything strange about the bag in an exit questionnaire. Only one of the 60 participants reported that they did, saying she noticed there were some marbles on the other side of some fabric which she could not grab.

[5]and a weak sense of morality in general.

bag, but could not see their partner's draws. On each trial, the experimenter held out the bag and participant 1 drew a marble from the bag, looked at it, and put it back in the bag. The experimenter then moved to participant 2 for them to do the same (see Figure 4.3). While the experimenter walked from one participant to the other, she flipped the switch on the handle to open the appropriate compartment and shook the marbles in the bag (to mask any sound of the switch and to simulate mixing the marbles so participants would think they had an independent draw from their partner). Participants took turns making draws in this way until each participant had seen 10 draws. A second experimenter seated off to the side wrote down the result of each draw to have a record of what participants actually saw, in case there were any errors in delivering the pre-generated sequences.[6] The writing was inaudible and kept in pen on a padded clipboard.

In the production phase, each participant was given one white, unlined index card and they were both asked to write down a marble color. They were told that the only goal in this part of the experiment is to write the same marble color that their parter writes on their paper. And if they both write the same marble color, they will receive a small monetary bonus. When both participants had written down their response, the cards were collected and placed face down on the ground. No feedback was given. Each participant was given another index card and the same instructions were repeated. Production trials were repeated until each participant had written down 10 marble colors. Response time was unconstrained and most participants took about 5 seconds per production trial. Participants wrote with graphite pencils or black pens to avoid giving them red or blue color cues. Participants then completed an exit questionnaire, looked at the outcome of their coordination attempts, and were debriefed. Each participant received a £1 bonus and was told it was "for the matches they got correct"[7], so that all participants would be paid equally for their time.

[6] All pre-generated sequences were delivered correctly.

[7] All pairs achieved at least one match.

Figure 4.3: Photos illustrating the observation phase procedure of Experiment 5, where a pair of participants take turns drawing marbles from the same bag. Top: participant one draws a marble and puts it back into the bag. Middle: the experimenter moves to the next participant, flipping the switch out of participants' line of site while "mixing" the marbles with her left hand. Bottom: participant two draws a marble and puts it back into the bag.

Conditions

Two different observation ratios were tested. 30 participants observed a 5:5 ratio and 30 participants observed a 7:3 ratio. Participants received the same observation ratio and counterbalance condition as their partner[8]. Participant behavior in this task is independent of the set of draws that their partner saw. The only reason that pairs were placed in the same condition was to make for a more intuitive debriefing.

4.2.2 Results

Coordination index

First we will discuss the ability of participants to coordinate in this game. Following Mehta et al. (1994a) and Bardsley and Mehta (2010), we calculate the coordination index, which is a summary statistic of how well participants are able to coordinate in this task. Rather than reporting the number of coordinations actually achieved per pair, this reports the probability that coordination will be achieved between any two participants in our participant pool, chosen at random without replacement. The coordination index is:

$$p_{coord} = \sum_{i=1}^{k} \frac{n_i(n_i - 1)}{N(N-1)} \tag{4.1}$$

where N is the total number of participants and n_i is the number of participants that gave each response $1...k$. This metric ranges from 0 (where no coordination is achieved) to 1 (where coordination is always achieved). For example, the coordination index of Mehta et al. (1994a)'s Heads and Tails result (where 87 participants responded *heads* and 13 responded *tails*) is $p_{coord} = 0.77$. Because this index returns the probability that two individuals drawn from the population *without replacement* will coordinate, the true chance level will be slightly lower than $p_{coord} = 0.5$. In the following sections, I will present the coordination indices per observation ratio condition. The chance-level index for the 30 participants in one observation ratio condition is $p_{coord} = 0.4915$.

[8]There was one exception due to odd numbers in the counterbalance condition: for one pair in the 7:3 condition, one partner saw 7 blues and 3 reds, and the other saw 7 reds and 3 blues. Again, participant behavior in this task is independent of their partner's behavior, so the type of data collected from this pair will be no different than the type of data collected for the other pairs.

One-shot coordination

Because participants were not aware that there would be multiple production trials, the first production trial constitutes a true one-shot game. This section presents the results of the first production trial only.

Color as a focal point. To determine whether or not participants successfully exploited color for coordination, I computed the coordination index where n_i encodes the color identity of participants' one-shot response. This tells us how well participants were able to coordinate by biasing their productions toward one of the colors. The results show that color was not successfully exploited for coordination. In the 5:5 condition, 14 participants produced the blue marble and 16 produced the red ($p_{coord} = 0.485, p = 0.86$). The p-value that I have chosen to report with the p_{coord} index is that of a two-tailed, exact binomial test. This gives the probability of obtaining a particular split of the two marble types (here, 14 blue and 16 red) if participants were randomly choosing each type with a probability of 0.5. If $p \leq 0.05$, this means that participants were coordinating on one of the variants significantly better than chance. In the 7:3 condition, 13 participants produced the blue marble and 17 produced the red ($p_{coord} = 0.492, p = 0.59$). Participants perform at chance levels in both observation ratio conditions: neither the red nor blue color identity was successfully exploited for coordination.

Frequency as a focal point. When we look at participants' coordination success based on the observation frequency for their response, participants perform significantly better than chance ($p_{coord} = 0.76, p < .001$). Here, the coordination index was computed so n_i encodes participants' response options as the *majority* or the *minority* marble. Participants in the 7:3 condition were the only ones who observed a majority marble. Of the 30 participants in this condition, 26 of them produced the majority marble in their first production trial. But perhaps this high number of majority productions is because participants were ignoring the coordination aspect of this task and just probability matching on their first production, as in Experiment 1. To assess this possibility, the p-values were re-computed with an exact binomial two-tailed test where the probability of the two responses is not 0.5, but equal to the observation ratio, 0.7. Even using a 0.7 binomial baseline, production of the majority marble is significantly higher than chance ($p = 0.047$). But as we saw in Experiment 1, binomial probability matching is not the appropriate baseline here. If participants were probability matching in this task, the variance would be restricted compared to that of the binomial and that means the p-value here should be lower than it is under binomial

sampling. Therefore, we can safely say that participants are not just probability matching in this task. Rather, participants are exploiting the common ground frequency information they share with their partner to succeed in this coordination task. Furthermore, it is clear that participants' overproduction of the majority variant is specifically due to a coordination attempt, and not some kind of general production bias, because participants in Experiment 1 who also saw the 7:3 observation did not produce the majority marble on their first production trial more than chance with the 0.7 binomial baseline ($p_{coord} = 0.58, p = 1$ in an exact binomial test with 32 trials and 23 majority marble productions)[9].

Coordination across all production trials

Although one production trial is enough to analyze participants' use of color and frequency as focal points, this section assesses how well participants were able to coordinate in this entire task by pooling all 10 production trials together, yielding 300 data points per observation ratio condition.

Color as a focal point. In the 5:5 condition, participants successfully coordinated on the blue marble, producing blue on 169 trials and red on 131 trials ($p_{coord} = 0.51, p = 0.03$). In the 7:3 condition, participants did not successfully coordinate on the basis of color, producing blue on 139 trials, red on 160 trials, and green on 1 trial ($p_{coord} = 0.497, p = 0.25$). Here, the coordination index was computed for all three variants produced (blue, red, and green) but the green trial was discarded for the binomial test. In the last section, we saw that color was not successfully exploited on the first production trial in either condition. However, across all 10 production trials color *was* successfully exploited by participants in the 5:5 condition, but not in the 7:3 condition. There are two possible explanations for the difference in 5:5 results. First, the one-shot analysis used 30 data points and the pooled analysis used 300 data points. It is possible that participants were trying to coordinate on blue in the first trial, but the effect was too small to detect in the small sample size. The second, and perhaps more interesting option, is that participants may have started trying to coordinate on blue sometime after the first production trial. There is evidence in repeated games without feedback, such as this one, that learning can take place and player behavior can become more rational, adjusting in the direction of Nash equilibria over time (Weber, 2003; Rick and Weber, 2010). Therefore, it could be the case that over the course of the production phase, participants figured out how to

[9]This p-value = 1 because 23/32 is as close to 0.7 as possible for 32 draws. These 32 participants *exactly* probability matched on their first draw at the population level.

coordinate on the basis of color saliency. However, I have no explanation for why blue was chosen over red. As discussed in Section 4.1.2, the literature shows that red and blue are both strong focal colors. Although the present result shows that blue was exploited more than red, Scott-Phillips et al. (2009) found red to be exploited more than blue. It is likely that these colors are nearly tied for focal strength and whichever one wins out in any given experiment may depend on the specifics of the experimental design and the exact shade of the red and blue stimuli.

Frequency as a focal point. In the 7:3 condition, participants successfully coordinated on the majority marble, producing it 193 times and producing the minority marble 107 times ($p_{coord} = 0.54, p < .001$ with a 0.5 binomial baseline). The 0.7 binomial baseline also yields a significant difference, showing that participants are not simply probability matching in this task ($p = 0.032$ with a 0.7 binomial baseline).

Regularization profiles: the entropy of variants

In this section, I present the distribution of ratios that participants produced and then assess participant regularization behavior as we did in Experiment 2, by calculating the change in entropy between each participant's observation and production ratio. I will also make some comparisons to other regularization profiles from Experiments 1 and 2. To understand how coordination behavior elicits regularization, above and beyond that obtained in the same frequency learning task, I will compare the results of Experiment 5 to Experiment 1. And to shed some light on the similarity between non-linguistic coordination and linguistic regularization, I will also compare Experiment 5 to the one-item word frequency learning in *words1*, from Experiment 2. Both comparisons will be restricted to the data obtained from the 5:5 and 7:3 observation ratios, because these were the only data collected in Experiment 5.

Figure 4.4 shows the distribution of ratios that participants produced in Experiment 5, per observation ratio, and re-plots the data from Experiment 1 and 2 for comparison. First let's look at the regularization profiles for the coordination task. In the 5:5 condition we see that most participants also respond with a 5:5 ratio, and only a few participants fully regularize by producing all red marbles or all blue marbles. In the 7:3 condition, more participants produce fully regular sequences, fewer participants produce the 5:5 ratio, and not many probability match by producing a 7:3 ratio. In fact, the mode of the response distribution is on the 6:4 ratio, biased toward the majority variant. Nearly all regularization

during coordination occurs in the direction of the majority marble.

In comparison to Experiment 1, we can see that the regularizing effect of coordination is not due to the basic frequency learning and production task at hand, because very little regularization behavior is obtained during probability matching. In comparison to Experiment 2 (*words1*), the regularization profile of word learning appears more similar to that of coordination. Upon visual inspection, it seems that participants learning about word frequencies also are most likely to produce a 5:5 ratio when they've observed a 5:5 ratio, and less likely to fully regularize. In the 7:3 condition, word learners also regularize by overproducing the majority variant. However, there are some differences as well. Coordinators seem less likely to probability match in the 7:3 condition than word learners do. To investigate this point I coded participants binarily as exact probability matchers or not. There were 12 perfect probability matchers in the 5:5 condition and 3 perfect probability matchers in the 7:3 condition. A Pearson's Chi-squared test (with Yates' continuity correction) showed a significant effect of observation ratio on the number of perfect probability matchers ($\chi^2(1) = 5.6889, p = 0.02$). Additionally, it seems that coordinators are more likely to fully regularize in the 7:3 condition (where they have a frequency-based focal point) than in the 5:5 condition (where they do not), whereas word learners seem equally likely to fully regularize, regardless of their observation ratio. However, this apparent difference in coordinators was not statistically significant. Participants were coded binarily as a full regularizer or not. There were 4 full regularizers in the 5:5 condition and 7 full regularizers in the 7:3 condition. A Pearson's Chi-squared test (with Yates' continuity correction) showed no significant difference in the number of full regularizers between observation ratio conditions in coordinators ($\chi^2(1) = 0.4453, p = 0.51$). The same was done for word learners, with 7 full regularizers in the 5:5 condition and 8 full regularizers in the 7:3 condition. This difference was not significant either ($\chi^2(1) = 0, p = 1$).

Figure 4.5 shows the overall amount of regularization elicited in Experiment 5, again with comparisons to the 5:5 and 7:3 data from Experiment 1 and Experiment 2 (*words1*). Here, we see that coordination elicits regularization behavior: the change in entropy in Experiment 5 is significantly lower than zero ($t(59) = -3.8887, p < .001$, one sample t-test, two-tailed). Coordination also elicits significantly more regularization behavior than the probability matching behavior of Experiment 1 ($t(76.97) = -3.1258, p = 0.003$). However, the amount of regularization elicited by coordination and word learning is not significantly different ($t(121.992) = 0.7644, p = 0.45$). In terms of regularization differences between observation ratios within an experiment, there are no significant differ-

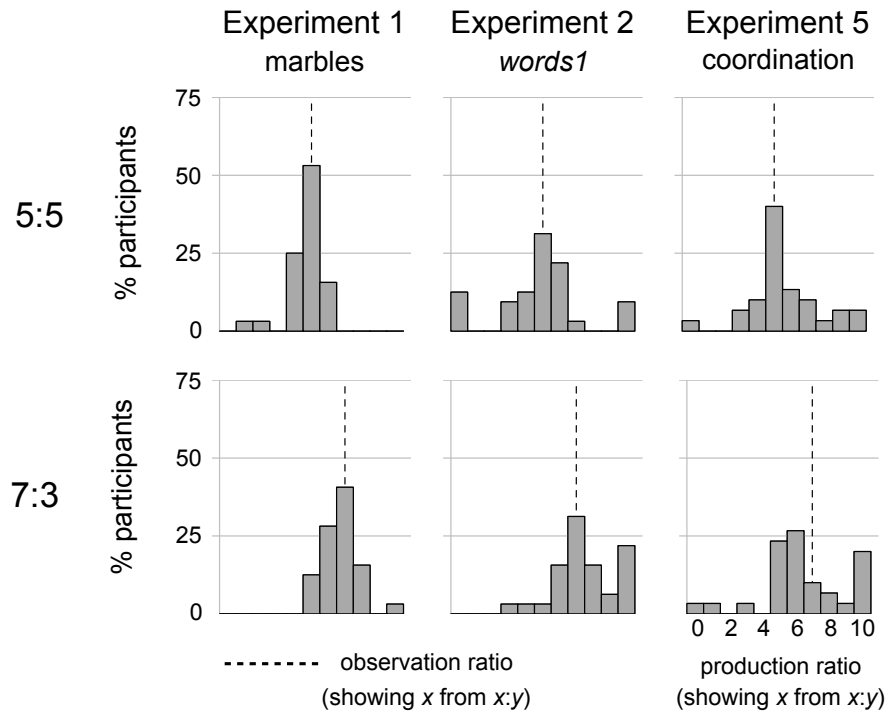


Figure 4.4: The distribution of ratios produced, per observation ratio (5:5 or 7:3) in Experiment 1, 2, and 5. The x-axis plots the number of times a participant produced the *observed majority* (x from the observation ratio (x:y) and the dotted line shows the frequency of the *observed majority*.

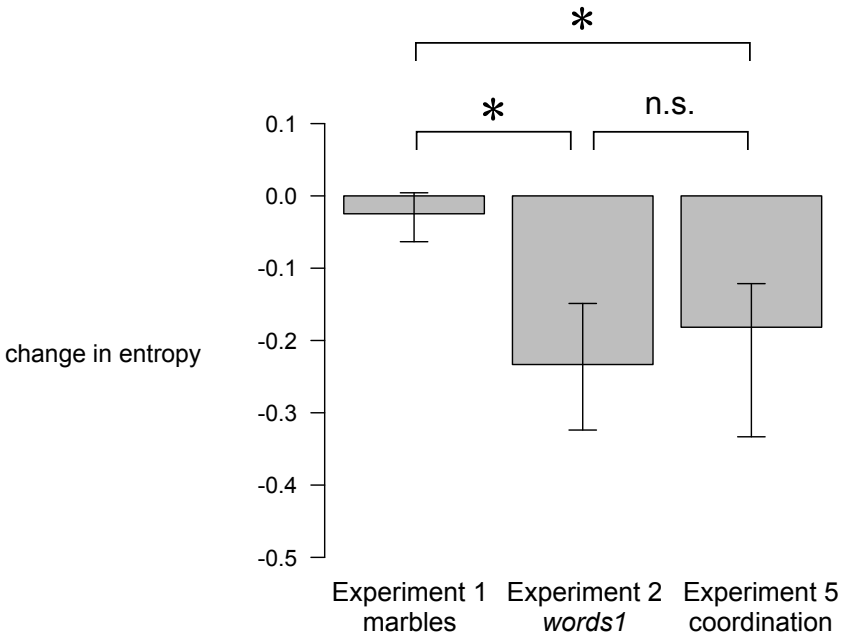


Figure 4.5: Mean entropy drop in Experiment 1, Experiment 2 (*words1* only), and Experiment 5. Each bar includes pooled data from the 5:5 and 7:3 observation ratios of each experiment.

ences (Experiment 1, 5:5 vs 7:3: $t(48.884) = -0.6686, p = 0.51$; Experiment 2 (*words1*), 5:5 vs 7:3: $t(61.576) = -0.2105, p = 0.83$; Experiment 5, 5:5 vs 7:3: $t(54.198) = 0.2219, p = 0.8252$). This is due to the large amounts of probability matching and production of 5:5 ratios obtained in these experiments, regardless of observation ratio.

Participant achievement of Nash equilibria

In a one-shot game, it is impossible to tell which strategy, if any, participants are employing when they produce either red or blue. But taking all production trials into account, participants can be classified according to the Nash equilibrium strategies. I will classify participants who only produced all reds or all blues as *pure strategy* players and participants who only produced a 5:5 ratio of reds and blues as *mixed strategy* players. This is a conservative definition of mixed strategy players because players do not know how many production trials there will be and could end up producing ratios off the 5:5 mark at the time the production phase cuts off. Of the 60 total participants, 10 adopted a *pure strategy* and 19 adopted a *mixed strategy* (Figure 4.6a).

Observation frequency modulation of Nash equilibria strategies

A different proportion of participants played the pure and mixed strategies in each observation ratio condition. In the 7:3 condition (Figure 4.6b), an equal number of participants played either strategy: 7 played a pure strategy and 7 played the mixed strategy. In the 5:5 condition (Figure 4.6c), 3 participants played a pure strategy and 12 played the mixed strategy. However, the observation ratio was not found to be a significant predictor of the strategy participants play ($\chi^2(1) = 1.710, p = 0.191$).

In the 7:3 condition, where frequency is a potential focal point, participants who play the pure strategy seem more likely to do so with the majority marble. Of the 7 participants who played the pure strategy, 6 played it with the majority marble and 1 played it with the minority marble. Despite the small data set, this difference is approaching significance ($\chi^2(1) = 3.5714, p = 0.059$), but we can not reject the null hypothesis that focal-frequency has no effect on which pure strategy participants decide to play. Because focal-frequency was shown to be a significant predictor of participant behavior on the first production trial, I believe that a larger sample of pure strategy players would show significantly more plays of the majority marble.

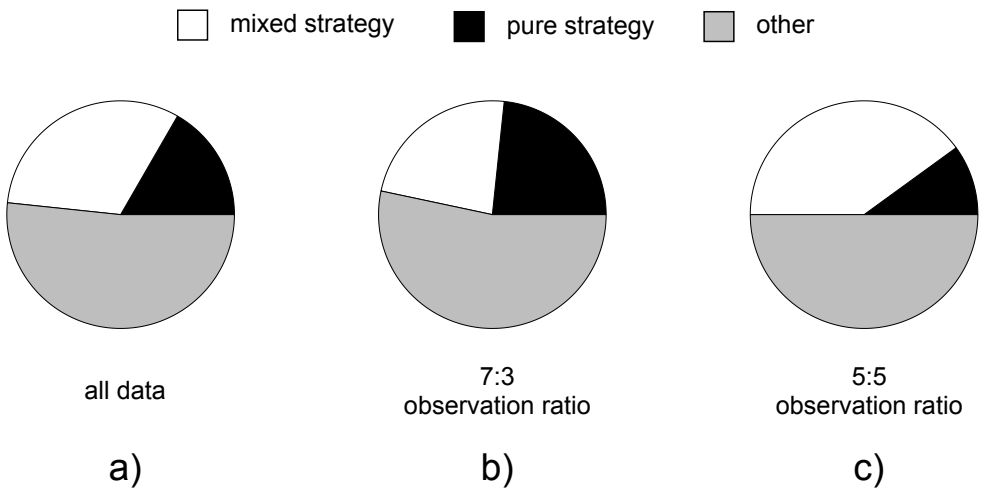


Figure 4.6: Breakdown of the strategies played by participants in Experiment 5 for a) all participants, b) participants in the 7:3 observation ratio condition, and c) participants in the 5:5 observation ratio condition.

Regularization: the conditional entropy of the sequences

Thus far in our analysis, we have been assessing regularization behavior in terms of the relative proportions of the variants, because this is how regularization has proceeded in all previous experiments. After the data was collected for Experiment 5, however, it became clear that many participants had produced some unusually structured sequences. Furthermore, the particular structures that they tended to produce constituted an alternative way to increase the predictability of their responses. Because this was a rather exploratory experiment and I am interested in all the ways that participants' tacit coordination behavior compares to behavior during linguistic frequency learning and production, I will include a post-hoc analysis of the sequence structures in this final section.

A structure-based coordination strategy would entail using some sort of rule system where a participant's choice on a trial is affected by their choice in a previous trial. For a first pass at this kind of structural analysis, we will determine whether a participant's choice on trial t is affected by their choice on the immediately preceding trial t-1. This will treat sequences as being generated by a first-order Markov model and ask whether there is good evidence that the transition probabilities between states are not equally weighted. Thus, our null hypothesis is a Markov model with 2 states, red and blue. From each state there are two edges, one going back to the same state and one going to the other state, and both of these edges are taken with equal probability. If we choose an initial state with equal probability and run this model for 10 time steps, a

length-10 sequence of marble draws will be generated. This model can generate 1024 sequences, which is the number of permutations of m elements in a length-n sequence, given by m^n. Each of these sequences has the same probability of being generated: $\frac{1}{1024}$. If we classify sequences according to their ratio (ex: 5:5), this random process is more likely to generate some ratios than others. The number of sequences in each ratio is $\frac{n!}{(n-k)!(k!)}$, which is the number of combinations of k reds and $(n-k)$ blues in a length-n sequence. The distribution of ratios that this process is expected to generate is as follows:

ratio of red:blue	0:10	1:9	2:8	7:3	4:6	5:5	6:4	7:3	8:2	9:1	10:0
combinations	1	10	45	120	210	252	210	120	45	10	1
probability	.001	.009	.044	.117	.205	.246	.205	.117	.044	.009	.001

These combinatorics result in different probabilities that each ratio will be produced by this random process. This probability distribution is given by the binomial equation (refer back to Equation 1.1), where $p = 0.5$. This constitutes the baseline we will be working with when determining whether participants are producing more structured sequences than expected by chance. I will come back to the importance of this baseline in a moment.

The measure of structure that we will use is the conditional entropy of a sequence. This tells us about the skew in weighting on the edges of the first-order Markov model by quantifying, in bits, how much information is needed to fully predict the next trial outcome, given that we know what the previous trial outcome was. An example calculation is given in Box 4.1.

The conditional entropy score is bounded by 0 and 1 for binary sequences when log base 2 is used in the calculation. Sequences in which each element is completely predictive of the element that follows it, such as 0101010101 or 111111111 have the minimum conditional entropy of 0 bits (and thus, are highly structured). Sequences in which each element carries no information about the element that follows it have the maximum conditional entropy of 1 bit, such as 01100 (and are not structured). For binary sequences of length 10, the maximum entropy is 0.98 bits and is achieved in sequences such as 0011010011, where the four possible transition types (00, 01, 10, 11) occur nearly equally[10]

[10]Only binary sequences where $\frac{length-1}{4}$ is an integer can have a maximum entropy of 1 bit.

Example 4.1. *Conditional entropy of a binary sequence*

The amount of structure in a binary sequence can be quantified by the conditional entropy of X (the set of elements at time t) given Y (the set of elements at time $t-1$):

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y) \tag{4.2}$$

Let's calculate the conditional entropy of an example binary sequence: 0110110110

Let y_0 be the number of 0s and y_1 be the number of 1s at time $t-1$. Let x_0 be the number of 0s and x_1 be the number of 1s at time t.

$p(y_0) = \frac{3}{9}$, $p(y_1) = \frac{6}{9}$
$p(x_0|y_0) = 0$, $p(x_1|y_0) = \frac{3}{3}$, $p(x_0|y_1) = \frac{3}{6}$, $p(x_1|y_1) = \frac{3}{6}$

$$H(X|Y) = \left(\frac{3}{9} \cdot (0 \log_2 0 + \frac{3}{3} \log_2 \frac{3}{3})\right) + \left(\frac{6}{9} \cdot (\frac{3}{6} \log_2 \frac{3}{6} + \frac{3}{6} \log_2 \frac{3}{6})\right)$$

$H(X|Y) = \frac{6}{9}$ bits

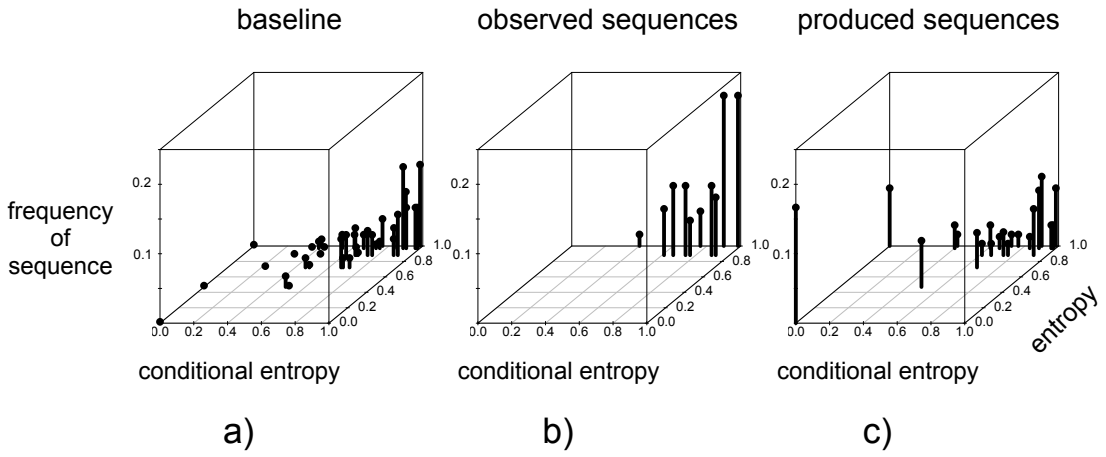


Figure 4.7: The entropy and conditional entropy of a) all binary sequences of length 10, which constitute the 1024 sequences participants can produce, b) all sequences used in the observation phase of Experiment 5, and c) all sequences that participants produced in Experiment 5. The skew of this distribution exemplifies regularization as a process that takes participant behavior into the low entropy regions of all possible behaviors.

Figure 4.7a plots each of the possible 1024 sequences in terms of their conditional entropy and entropy. If participants are producing sequences by randomly selecting between red and blue on each production trial, then the 60 sequences obtained in Experiment 5 should conform to a random sample from this baseline distribution. Figure 4.7b shows the location of the 60 observation sequences in this space, which constitute a random sample in the 5:5 ratio (where entropy = 0.88 bits) and 7:3 ratio (where entropy = 1 bit). Figure 4.7c shows the location of the 60 sequences that participants produced. These sequences are not likely to be a random sample from the baseline distribution in terms of entropy ($t(59) = -3.4583, p = 0.001$) or conditional entropy ($t(59) = -4.6118, p < .001$).[11] They are highly skewed toward predictable sequences with low entropy and low conditional entropy. The skew of this distribution exemplifies regularization as a process that takes participant behavior into the low entropy regions of all possible behaviors.

In particular, there are four sequences that account for the largest share of regularization. Two of these sequences are located in the bottom left corner of Figure 4.7c, where conditional entropy and entropy are both 0 bits. These sequences are 0000000000 and 1111111111, and correspond to the pure Nash

[11] I conducted a one-sample, two-tailed t-test comparing the entropy/conditional entropy of participant productions to the mean of the baseline distribution over entropy/conditional entropy values.

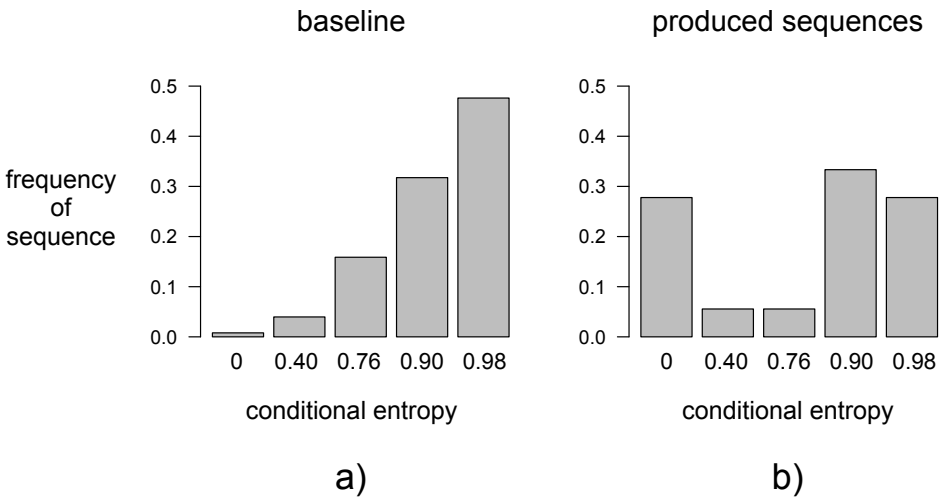


Figure 4.8: Distribution of conditional entropy scores for a) all possible sequences in a 5:5 ratio and b) all 5:5 sequences participants produced in Experiment 5.

equilibrium strategy. The other two sequences are located in the top left corner where conditional entropy is at its minimum, but entropy is at its maximum. These are the two alternating sequences: 0101010101 and 1010101010, which correspond to the mixed Nash equilibrium strategy. These salient sequences, which make up 0.4% of the sequences participants could have produced in this task, constitute 25% of participants' actual productions.

In Figure 4.7a, we can see that the entropy and conditional entropy of sequences are correlated. If something such as a regularization process is causing participants to overproduce one of the variants, leading to a drop in entropy, then the conditional entropy score of the sequence will also go down, even though there was no process at play that explicitly targeted the temporal structure of the sequence. Therefore, to be sure that the observed drop in conditional entropy is not due to a variant-based regularization process, we should look for drops in conditional entropy within a given production ratio, because all of these sequences will have the same entropy. Within-ratio drops in conditional entropy provide more definitive support for the presence of a regularization process that targets the temporal structure of participants' productions. I will restrict this within-ratio analysis to the 5:5 ratio, because this is the ratio with 1) the widest range of conditional entropy values and 2) the highest number of sequences, both theoretically and in the data obtained from participants.

Figure 4.8a shows the baseline distribution of conditional entropy scores across the 252 possible sequences in a 5:5 ratio and Figure 4.8b shows the distribution of conditional entropy scores for the 18 5:5 sequences that participants produced

in Experiment 5. The sequences that participants produced are not likely to be a random sample from this baseline distribution ($t(17) = -2.4982, p = 0.02$).[12] Participant productions are more structured, with a mean of 0.64 bits, whereas the baseline mean is 0.89 bits.

Are participants producing structured sequences because they are in a tacit coordination situation, or could it be that people always produce structured sequences regardless of the task at hand? It is well-known that humans have systematic biases in random sequence perception (Kahneman and Tversky, 1972; Hahn and Warren, 2009) and production (Tune, 1964). When people attempt to generate random sequences, their productions tend to be more rule-governed (Brugger, 1997) with repetitions of specific subsequences and too many alternations, resulting in lower average conditional entropy than truly random sequences (Baddeley, 1966). To determine whether it is the pressure to coordinate that causes participants to produce more structured sequences, I compared the conditional entropy scores of the 5:5 productions across all experimental conditions reported in this thesis. Figure 4.9 shows the mean values for Experiment 1, the four conditions in Experiment 2, Experiment 5, and the baseline distribution for a 5:5 ratio. All distributions, except for the words6 condition, are significantly unlikely to be a random sample from the 5:5 baseline distribution [13]. They all are more structured than the baseline, but Experiment 5 shows the highest degree of structure.

Of the data sets in Figure 4.9, Experiment 1 provides the closest coordination-less comparison to the task of Experiment 5. The observation phase of both of these experiments was essentially the same: participants observed 10 marble draws from a bag (with blue/orange marbles in Experiment 1 and blue/red marbles in Experiment 5). The production phases were also similar: participants produced a sequence of 10 marbles, one at a time, and they could not see a history of their past productions. However, Experiment 1 framed the production task by asking participants to indicate what several more random draws from this bag are likely to look like and Experiment 5 framed it as a tacit coordination game. Therefore, the amount of structure obtained in Experiment 1 should give us an idea of how biased participants' attempts of random sequence production are in this type of task (without a coordination component). The set

[12] As above, I conducted a one-sample, two-tailed t-test comparing the conditional entropy of participant productions to the mean of the baseline distribution over conditional entropy values, but only for sequences in the 5:5 ratios.

[13] Experiment 1: $\chi^2(4) = 87.8926, p < .001$. Experiment 2, *marbles1*: $\chi^2(4) = 14.9375, p = .005$, *marbles6*: $\chi^2(4) = 30.8059, p < .001$, *words1*: $\chi^2(4) = 30.8059, p < .001$, *words6*: $\chi^2(4) = 4.4324, p = .35$. Experiment 5: $\chi^2(4) = 167.967, p < .001$.

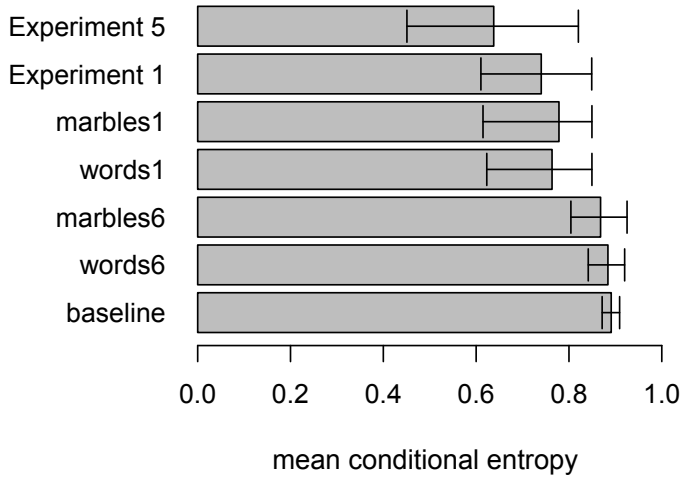


Figure 4.9: Mean conditional entropy of all 5:5 production sequences in Experiment 5, Experiment 1, and the four conditions in Experiment 2: *marbles1*, *words1*, *marbles6*, *words6*. The baseline mean is that of all possible 5:5 production sequences and represents the average conditional entropy that would be expected if participants were not biasing their productions toward structured sequences. Error bars are bootstrapped 95% confidence intervals (Efron, 1979).

of sequences produced in Experiment 5 are significantly more structured than those of Experiment 1 $\chi^2(4) = 9.598, p = 0.047$. This means that the pressure to coordinate causes participants to structure the sequence of their productions above and beyond their tendency to do so when attempting to produce random sequences.

4.2.3 Discussion

In this chapter we have touched on several aspects of non-linguistic coordination and regularization behavior. The most important result was that participants can coordinate in this task, and they do so by spontaneously exploiting the relative frequency of variants in an environment they share with their partner. This behavior only makes sense if participants were able to infer the relative frequency of marbles in the bag, on the basis of their observed sample from the bag, and infer that their partner was also likely to see a sample from the bag which was skewed toward the same marble color that their sample was skewed toward. In some ways, this is very similar to what takes place during coordination in language. There exists a veridical distribution over utterances in any given language, for a specified time frame, but every user of this language has only experienced a sample of these utterances. If two users of the language are to coordinate on the same

arbitrary signal for a meaning, they could exploit the relative frequency of the signals for that meaning to achieve coordination. This is exactly what language users seem to do when they are taught a variable language: they produce a more deterministic language by overproducing the signal that had the highest relative frequency in their sample.

In this experiment, participants preferred to coordinate on the basis of a variant's frequency, rather than its content: they did not exploit color-based focal points when the marbles could be distinguished on the basis of their relative frequency. This may be because the two colors that were provided as common ground (red and blue) are the two most preferred colors for coordination. If only one of these colors were used along with a dispreferred coordination color, such as beige, then it is possible that participants would have exploited color for successful coordination even if one of these colors occurred in the majority of observations. However, it may be the case that frequency information is always given preference. Through personal communication with Karen Schloss (at the Palmer Visual Perception and Aestheics Lab in UC Berkeley), higher-level cognitive mechanisms involved in frequency learning may be trumping people's basic color preferences. According to her, the blue and orange pair of marble colors used in Experiment 1 are the most and least preferred colors, respectively (according to an United States participant pool) and blue should have been over-produced in Experiment 1, but this was not the case. Instead, the frequency learning task in Experiment 1, which lead participants to reliably probability match, seems to have overridden participants' basic color preferences. In Experiment 5, the higher-level cognitive mechanisms involved in tacit coordination did lead participants to coordinate on the preferred color (blue), rather than probability match, but when frequency information was added on top of the same marble color information, participants abandoned the focal-color strategy and coordinated on the majority marble only. It is quite possible that coordinating individuals privilege frequency-based information over content of signals in all tacit coordination games, precisely because the solution to such games is arbitrary and agnostic to content. This would imply that, in the coordination game of natural language use, the frequency of linguistic features may play a larger role in language learning and production than their various content features do.

This experiment also demonstrated that the pressure to coordinate elicits regularization behavior, and the amount of regularization it elicits was not significantly different from the amount elicited by word learning. This suggests that the same mechanism may underlie regularization in language and in non-linguistic coordination. However, there were some differences when we took a more detailed

look at the regularization profiles. Although word learners and coordinators both fully regularize, some word learners also seemed to probability match, whereas coordinators do not: instead, some coordinators play a mixed strategy, producing 5:5 ratios.

Another difference unique to coordination is the amount of regularization obtained in terms of sequence structure. Coordinating individuals produced sequences with lower conditional entropies than participants in any other experiment in this thesis. This may be because whole sequences can also serve as focal points, and is evidenced by the fact that participants converged upon 4 specific sequences (all red, all blue, red-blue alternating, and blue-red alternating) with a probability much higher than chance. If these highly structured sequences do serve as focal points, then changing the production task to be more explicitly about whole-sequence matching may increase the percentage of participants that produce these four sequences. For example, instead of coordinating color by color in a series of one-shot games, participants could construct a sequence of 10 draws. This sequence would then be compared to their partner's and only if all marbles match up, would they receive a reward. I hypothesize that a production task such as this will lead to even more structure-based regularization (i.e. lower conditional entropy of the sequences participants produce). However, such structure-based coordination does not seem to have any parallels in the timecourse production of synonyms.

In summary, participants exploit frequency in a non-linguistic coordination game, often by playing the pure Nash equilibrium strategy on the majority variant, and this leads to regularization behavior. Participants do not regularize in a matched task without coordination, but do regularize to a similar extent in a word learning task, which is not an explicit coordination game. The similarities in regularization between coordination and word learning suggest a functional account of coordination as a mechanism that contributes to regularity in language. In this way, coordination may be a key component of the domain-specific regularization bias described in Experiment 2, if a linguistic framing of a frequency learning task is enough to trigger coordination strategies in participants. Or, it could constitute yet another contributor to the overall regularity observed in language, in addition to the domain-specific and domain-general drivers described in Experiment 2.

Chapter 5

Inductive biases and Bayesian model fitting

In the previous chapters, we have been looking in detail at human frequency learning behavior and experimentally determining what aspects of the learning context affect that behavior. Most notably, it is the type of stimuli (linguistic or non-linguistic), the number of frequencies a participant must track at one time (for one versus six items), and the goal of the task (coordination or imitation[1]) that affect the amount of variation participants maintain or eliminate in their productions. These different learning contexts modulate the cognitive system by triggering different learning mechanisms, each of which may have its own bias. Because word learners, concurrent frequency learners, and coordinators demonstrated a consistent bias toward producing regular mappings, this is evidence that these three learning contexts engage biased cognitive processes.

In Chapter 1, I discussed culture as having two distinct phases of its life cycle: an internal phase, where it exists in the neuronal architecture of a cognitive system, and an external phase, where it exists as behavioral and concrete artifacts. The data obtained from behavioral experiments only allows us direct access to this external phase. However, behavioral data provides clues to the internal states when analyzed, and statistical tests help determine whether behaviors across experimental conditions are the result of different cognitive mechanisms or not. In this chapter, I sharpen the focus on the internal phase of culture's life cycle and ask what cognition-internal forces could be producing the various degrees of biased behavior that participants demonstrate. Short of doing neuroscience, cognitive modeling allows us to assess hypotheses about what is going on inside

[1]Here I'm referring to Experiments 1 though 4 loosely as imitation tasks, because the only goal was to produce draws that are likely to come from the bag, or to name the object as you observed it being named.

people's heads as they produce the behavior that they do. Improving our picture of these cognition-internal mechanisms helps us sketch the environment through which culture and languages replicate. Just as we might want to find out what caused some change in a population of organisms, such as why some particular trait exists with such a high frequency (is it because of some selective pressure, or could it just be due to drift?), we want to know what forces give rise to the particular distribution of variants in a population of cultural artifacts. This is done by positing different models of these processes and evaluating which model best accounts for the data at hand.

The major force that shapes culture is the inductive bias of the cognitive agents that transmit it. As discussed in Chapter 1, cultural transmission is a reverse engineering process where agents generate behaviors in response to other, observed behaviors. In the case of language learning, this may involve an inference of the grammar underlying observed utterances, and this inferred grammar would then be used to generate more utterances. In the case of frequency learning, agents see events occurring with specific frequencies, such as two marble colors being drawn from a bag, and must infer the relative proportion of marbles in the bag to predict what further draws from that bag might look like. If a learner's predictions consistently deviate from the observed frequencies, then an inductive bias must be at play.

A special class of cognitive models, Bayesian models, allow for the explicit quantification of inductive biases that learners employ when making inferences about the processes that generate observed data. In this chapter, I apply a Bayesian model of frequency learning developed by Reali and Griffiths (2009) to quantify participants' inductive biases during frequency learning to the five experiments in this thesis. In this model, the inductive bias is represented as a prior probability distribution over all possible observation ratios, such that certain ratios may be a priori more probable. When participants are viewing marble draws from containers, they may think 50/50 ratios are more plausible, but when participants are viewing words being used to name objects, they may think deterministic mappings are more plausible. Learners then combine their a priori biases with the data they observed to arrive at an estimate of the relative frequency of the marbles in the bag, or the relative frequency of two synonyms in a language. When participants are asked to imitate their observations, they may use this estimate during production. In this way, participants' productions may be biased due to their inductive bias in frequency estimation.

In the following sections, I will describe this model in more detail and then quantify participants' frequency learning biases by fitting this model to my em-

pirical data. Essentially, the model fitting procedure asks: what kind of bias would a Bayesian learner need to have to produce the data that participants produced? Although I do not assume that participants are perfect Bayesian rational learners, the model fits may be a fair approximation of participants' inductive biases. Therefore, I will assess how well this model fits the data and detail its capabilities and limitations in describing human probability matching and regularization behavior. Finally, I will assess the insights gained and suggest some improvements to this modeling framework.

5.1 A model of frequency estimation

This section describes the beta-binomial Bayesian model of frequency estimation as developed by Reali and Griffiths (2009). Assume that a set of variants exist in the world with a particular relative proportion to one another (such as a 70/30 ratio of marbles in a bag) and a learner observes N instances of those variants (such as N draws from that bag). Here, we will only consider sets of two variants, but this model generalizes to sets of any size. Let x denote the number of occurrences of *variant x* and let y denote the number of occurrences of *variant y*, which equals $N - x$. The learner's estimate of the true proportion of each variant is the world is denoted by θ_x and θ_y, respectively. The problem that the learner is faced with is estimating θ_x and θ_y from their observations x and y. Since θ_y follows directly from θ_x, and y follows directly from x, we can solve this estimation problem in terms of x and θ_x only. From here onward, an un-subscripted θ will refer to the estimate of the proportion of *variant x*.

This frequency estimation problem will be solved through Bayesian inference. A Bayesian rational learner uses Bayes' rule (Equation 5.1) to infer what hypothesis (θ) generated the observed data (x).

$$P(\theta|x) \propto P(x|\theta)P(\theta) \tag{5.1}$$

Bayes' rule combines the prior probability over all hypotheses, $P(\theta)$, with the likelihood of the data under each hypothesis, $P(x|\theta)$, to arrive at a posterior probability of each hypothesis given the data, $P(\theta|x)$. The learner then uses these posterior probabilities to inform their choice of a hypothesis.

In the beta-binomial implementation of Bayes' rule, the prior follows a beta distribution and the likelihood follows a binomial distribution. Here, the prior is a beta distribution over all hypotheses (Equation 5.2), where $B(\cdot, \cdot)$ is the beta function (Boas, 1983).

$$P(\theta|\beta_1, \beta_2) \sim Beta(\beta_1, \beta_2) = \frac{1}{B(\beta_1, \beta_2)}\theta^{\beta_1 - 1}(1-\theta)^{\beta_2 - 1} \tag{5.2}$$

For the purpose of exploring inductive biases related to regularization and probability matching, we will only consider a subset of priors which follow a symmetrical beta distribution (as in Reali and Griffiths (2009)). This reduces the beta distribution to one parameter, $\frac{\alpha}{2}$, such that:

$$P(\theta|\tfrac{\alpha}{2}, \tfrac{\alpha}{2}) \sim Beta(\tfrac{\alpha}{2}, \tfrac{\alpha}{2}) = \frac{1}{B(\frac{\alpha}{2}, \frac{\alpha}{2})}\theta^{\frac{\alpha}{2} - 1}(1-\theta)^{\frac{\alpha}{2} - 1} \tag{5.3}$$

The $\frac{\alpha}{2}$ parameter captures how regular or variable learners expect their observations to be.[2] Figure 5.1 shows some example priors from the range of symmetrical priors achievable by this model. When $\frac{\alpha}{2} < 1$, learners have a regularization bias and their prior is symmetrically peaked at $\theta = 0$ and $\theta = 1$. When $\frac{\alpha}{2} > 1$, learners have a variability bias and their prior is peaked at $\theta = 0.5$. When $\frac{\alpha}{2} = 1$, learners are unbiased and their prior is flat, meaning they expect all hypotheses to be equally likely. The prior distribution gradually becomes peakier (unimodally) as $\frac{\alpha}{2}$ gets larger and peakier (bimodally) as $\frac{\alpha}{2}$ gets smaller than 1.

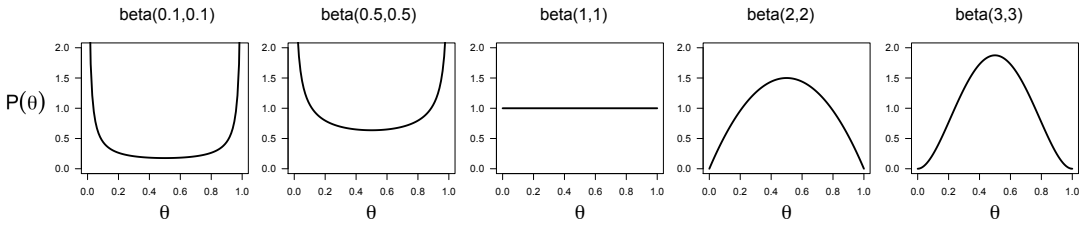


Figure 5.1: Example symmetrical prior beta distributions.

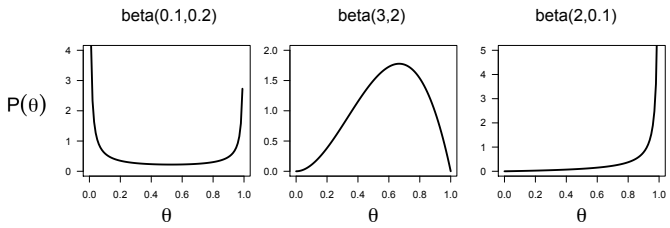


Figure 5.2: Example asymmetrical prior beta distributions.

[2] I kept the $\frac{\alpha}{2}$ notation of Reali and Griffiths (2009) to prevent confusion in cross-referencing. This is a notational hang-over from the Dirichlet-multinomial version of this model where the number of variants, k, is larger than two and the prior parameter is $\frac{\alpha}{k}$. For example, if participants were estimating the relative proportion of three marbles in a bag, I would use a Dirichlet-multinomial Bayesian model with a prior parameter $\frac{\alpha}{3}$.

All asymmetrical priors are excluded from this model. Some examples of asymmetrical priors are shown in Figure 5.2. By definition, all asymmetrical priors incorporate a direct bias toward *variant x* or *variant y*. Because there was no behavioral evidence of direct biases for the particular marble or word stimuli used, these priors will not be considered.

The next component of Bayes' rule is the likelihood function, $P(x|\theta)$. This assigns a probability of observing each value of x under each value of θ. For our case of two variants, the likelihood function is defined by a binomial distribution:

$$P(x|\theta, N) \sim Binomial(\theta, N) = \binom{N}{x}\theta^x(1-\theta)^{N-x} \tag{5.4}$$

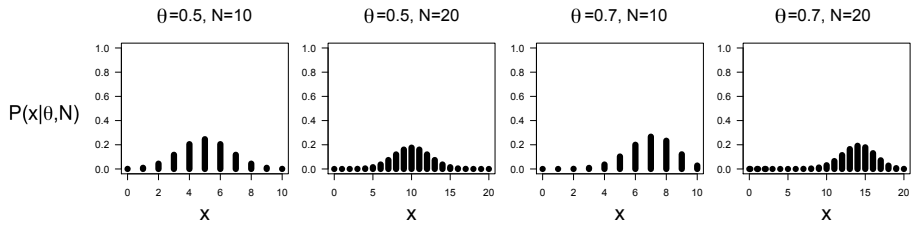


Figure 5.3: Example likelihood binomial distributions.

Figure 5.3 shows some examples of binomial distributions for different values of θ and N. Under a binomial likelihood function, each estimate always assigns the highest probability to its equivalent value of x. For example, if a bag contains a 50/50 ratio of blue to red marbles, the most likely outcome in a series of 10 draws is 5 blues and 5 reds ($\theta = 0.5$ and $x = 5$). Likewise, if a bag contains a 70/30 ratio, then the most likely outcome is 7 blues and 3 reds ($\theta = 0.7$ and $x = 7$).

The last component of Bayes' rule is the posterior probability over hypotheses, $P(\theta|x)$. Because the beta prior is conjugate to the binomial likelihood (Raiffa and Schlaifer, 1961; Gelman et al., 2013), the posterior will also be a beta distribution, and can be expressed simply as:

$$Beta(\beta_1 + x, \beta_2 + N - x) \tag{5.5}$$

Where, in the case of the symmetrical prior model:

$$P(\theta|x, N, \tfrac{\alpha}{2}) \sim Beta(\tfrac{\alpha}{2} + x, \tfrac{\alpha}{2} + N - x) = \theta^{\frac{\alpha}{2}-1+x}(1-\theta)^{\frac{\alpha}{2}-1+N-x} \tag{5.6}$$

This distribution provides learners with the posterior probability of each hypothesis and concludes the process of Bayesian inference. Some examples of posterior distributions for N = 10 and different values of x and $\frac{\alpha}{2}$ are shown in Figure 5.4.

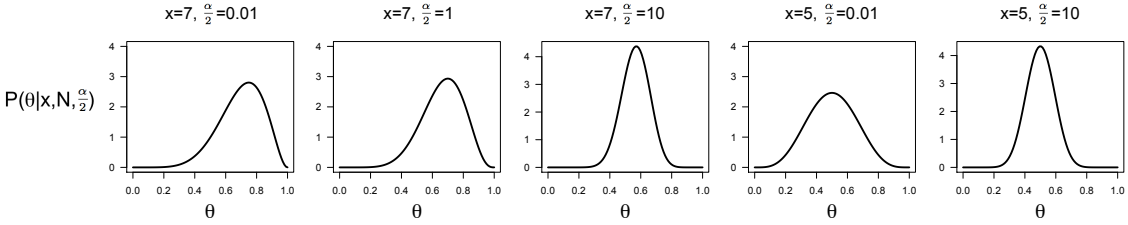


Figure 5.4: Example posterior beta distributions.

But how does a learner use these posterior probabilities to generate new data? In the frequency learning Experiment 2, participants observed 10 marble draws or naming events that constituted their input data, x. They were then asked to produce several more draws or naming events that may be likely to occur. Participants must have formed some estimate of the probabilities of the variants associated with each referent and generated a new set of variants from this estimate somehow. There are a variety of ways to implement hypothesis selection and data generation for a Bayesian rational learner. Three reasonable hypothesis choice strategies are:

1. *maximizing*: learners choose the maximum a posteriori hypothesis (MAP), which corresponds to the mode of the posterior distribution: $\frac{\frac{\alpha}{2}+x-1}{\alpha+N-2}$
2. *averaging*: learners choose the hypothesis at the posterior mean: $\frac{\frac{\alpha}{2}+x}{\alpha+N}$
3. *sampling*: learners choose a hypothesis by randomly sampling it from the posterior distribution, according to the posterior probability of each hypothesis.

After a hypothesis has been chosen by one of the above methods, data is generated from the hypothesis according to its likelihood under that hypothesis. Since the likelihood function of this beta-binomial Bayesian model is binomial, data production will also follow a binomial distribution. This is equivalent to generating data by randomly drawing marbles/words from the hypothesized proportion, with replacement. Rational learners must have complete knowledge of how data are generated from hypotheses (i.e. what the data likelihoods are). In a cultural transmission situation, the production likelihoods of one agent will be the likelihoods the next agent should use during induction. When all agents in the population are assumed to have identical learning algorithms, as most research

in cultural evolution assumes and will be assumed here, this means that the induction and production likelihoods should be identical for one learner.

So now we have a complete model that takes a learner from a specific type of input data (x) to a specific type of output data (x'). Reali and Griffiths (2009, p.321) provide equations for calculating the probability of producing any value of x' from any value of x for the sampler model (Equation 5.7) and the averager model (Equation 5.8).[3]

$$P(x'|x) = \binom{N}{x'} \frac{B(x + x' + \frac{\alpha}{2}, 2N - x - x' + \frac{\alpha}{2})}{B(x + \frac{\alpha}{2}, N - x + \frac{\alpha}{2})} \tag{5.7}$$

$$P(x'|x) = \binom{N}{x'} \left(\frac{\frac{\alpha}{2} + x}{\alpha + N}\right)^{x'} \left(1 - \left(\frac{\frac{\alpha}{2} + x}{\alpha + N}\right)\right)^{N - x'} \tag{5.8}$$

The equation for the maximizer model is obtained by replacing the term for the posterior mean in Equation 5.8 with the posterior mode, as follows:

$$P(x'|x) = \binom{N}{x'} \left(\frac{\frac{\alpha}{2} + x - 1}{\alpha + N - 2}\right)^{x'} \left(1 - \left(\frac{\frac{\alpha}{2} + x - 1}{\alpha + N - 2}\right)\right)^{N - x'} \tag{5.9}$$

These equations put the model into a format that we can use to fit to our experimental data, because our data is in the format of input-output (x to x') transitions and these equations assign a probability to each of these transitions. The model fitting task at hand is to determine which prior bias and hypothesis choice strategy assigns the greatest probability to participant behavior, and thus, best explains participant behavior.

5.2 Model fitting procedure

This section explains how the sampler, averager, and maximizer models are fit to participant data by maximum likelihood estimation of the prior parameter $\frac{\alpha}{2}$.

The best-fit model is the one that assigns the highest probability to the set of x to x' transitions produced by participants. Varying the model's parameters varies the model's fit. We want to determine which combination of parameters gives the model the best fit (i.e. explains participant behavior best). In our

[3]In this paper they refer to the averager as a MAP learner, although their equations are explicit about the use of the posterior mean for hypothesis selection. To avoid any confusion in terminology, I will stick to my labels of sampler (consistent with their terminology), averager (which produces data distributions that correspond to the posterior predictive distribution), and maximizer (which I use as a synonym for MAP).

experimental setup, the parameters N, x, and x' are given: these are the number of marble draws or naming events that participants observe per referent ($N = 10$), the frequency with which they observed *variant* x per referent ($x = 0, 1, 2, 3, 4,$ or 5, depending on the observation ratio), and the frequency with which they produced *variant* x per referent ($x' = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9,$ or 10). The only parameter that is not known is the one corresponding to participants' prior bias, $\frac{\alpha}{2}$. The following describes the maximum likelihood estimation procedure used to determine the best-fit value of $\frac{\alpha}{2}$ for a given model (sampler, averager, or maximizer).

First, a candidate value of $\frac{\alpha}{2}$ was plugged into the model and then the probability of each (x, x') transition in the data set was calculated under that model (via its corresponding equation, 5.7, 5.8, or 5.9). Next, the natural log of each probability were summed to yield the log likelihood score for this data under the $\frac{\alpha}{2}$ candidate in question. For example, take this hypothetical data set with 5 data points: $(x, x') = \{$(5,4),(5,5),(5,5),(5,7),(5,8)$\}$, where $N = 10$, and the candidate $\frac{\alpha}{2} = 1$. The probability of each data point under the sampler model is {0.165, 0.180, 0.180, 0.126, 0.077}. The natural log of these probabilities are {-1.80, -1.71, -1.71, -2.07, -2.56} and their sum is -9.85. This is the log likelihood score for this data set given a sampler model where $\frac{\alpha}{2} = 1$. This calculation was repeated for a discrete set of candidate $\frac{\alpha}{2}$ values in the range of $0 < \frac{\alpha}{2} \leq 100$, in 0.001 increments. The maximum likelihood estimate for the best-fit value of $\frac{\alpha}{2}$ is the one with the log likelihood score closest to zero. All scores in this range were unimodal, with one global maximum (refer ahead to 5.6).

To find out what percentage of the data a model explains, we can work out the prediction rate from the log likelihood of the model:

$$Prediction\ rate = exp\left(\frac{log\ likelihood}{number\ data\ points}\right) \tag{5.10}$$

The $exp(\cdot)$ function is used when the natural logarithm (log base e) is used in the log likelihood calculations. Otherwise, the argument of $exp(\cdot)$ should be multiplied by the natural log of the base used. The prediction rate for our example data set above is $exp(\frac{-9.85}{5}) = 14\%$. This means that the model, on average[4], will correctly predict each data point 14% of the time. This is equivalent to saying that the model accurately predicts 14% of participant behavior.

[4]Because the averaging was done in log space, this average is the geometric mean, which is the appropriate mean for comparing rates of change.

All of model fitting in this chapter will be carried out on data sets that exclude the 10:0 observation ratio because the maximizer model is undefined for $y = 0$ when $\frac{\alpha}{2} \leq 1$ and can not assign probabilities to participant production in response to the 10:0 observation ratio. This is known as the sparse data problem (e.g. Murphy, 2001), which is common in maximum likelihood estimation for rare or non-existent data points, and applies to the 10:0 observation ratio data because input data points for *variant y* are non-existent. There are various methods for dealing with this problem, but I will deal with it by just excluding the 10:0 observation ratio data from analysis. Additionally, the 10:0 observation ratio provides the least informative data for discriminating between different models because perfect probability matchers and perfect regularizers will respond identically to a 10:0 observation ratio, by producing only 10:0 observation ratios.

As a terminological note, I will refer to the best-fit priors as weak, medium, or strong biases, depending on their values of $\frac{\alpha}{2}$. Following Perfors (2012), I will refer to priors in the order of magnitude of tenths ($0.1 \leq \frac{\alpha}{2} < 1$) as a weak regularization bias, hundredths ($0.01 \leq \frac{\alpha}{2} < 0.1$) as a medium regularization bias, and thousands and below ($0 \leq \frac{\alpha}{2} < 0.01$) as a strong regularization bias. Likewise, I will apply this same classification to the variability bias and refer to priors in the order of magnitude of ones ($1 < \frac{\alpha}{2} \leq 10$) as a weak variability bias, tens ($10 < \frac{\alpha}{2} \leq 100$) as medium, and hundreds and above ($100 < \frac{\alpha}{2} \leq \infty$) as strong.

5.3 A note on comparing model fits

Normally, several models are fit to one data set to determine which model best explains the data at hand. In the following sections, I will be fitting the same set of models to all of the different data sets collected in this thesis. In many cases, the number of data points collected differs per experiment and per experimental condition. This means that the raw log likelihood scores cannot be directly compared across data sets. Log likelihood scores will generally be much lower for data sets that contain more data points because each observation receives a log likelihood under the model, and more observations mean more log likelihood scores are summed (remember, these scores are all negative). Instead, prediction rates should be used to compare fits across data sets. As seen in Equation 5.11, the prediction rate is based on the average log likelihood per data point, and this is what makes fair comparisons possible.

There remains one further issue in comparing model fits across data sets. Different data sets can contain different amounts of variability and prediction rates will necessarily be lower for data sets that are more variable. Only deterministic processes can be predicted with 100% accuracy and the ceiling on prediction rates is relative to the entropy of the data set as a whole. For example, Figure 5.5 reprints the distribution of production ratios from the 5:5 observation ratio in *marbles1* (left) and *marbles6* (right).

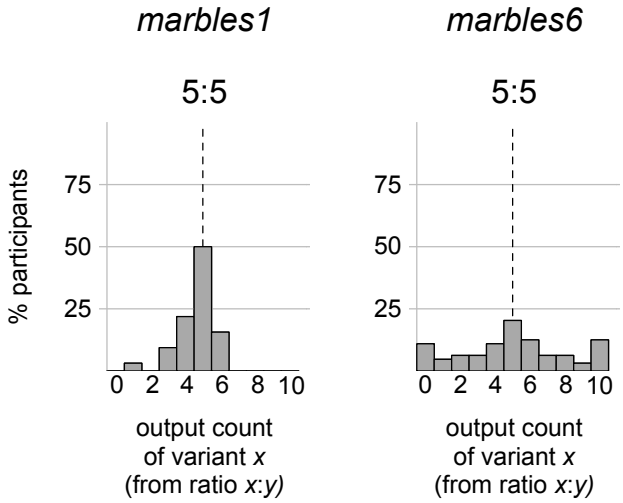


Figure 5.5: Reprints of the distribution of production ratios from the 5:5 observation ratio in *marbles1* (left) and *marbles6* (right). The entropy of each distribution is 1.87 bits for *marbles1* and 3.28 bits for *marbles6*.

At the *population* level, probability matching profiles are more deterministic than regularization profiles: the entropy of the *population* of responses to the 5:5 observation ratio in *marbles1* is 1.87 bits and in *marbles6* it is 3.28 bits. This means that any participant's response in *marbles1* is more predictable than in *marbles6* and thus, will have a higher ceiling on prediction rate than *marbles6*. The ceiling on the prediction rate can be calculated by fitting the data to itself, because the closest description of any data set is itself. This yields a ceiling of 27% for the example distribution from *marbles1* and a ceiling of 10% for the example distribution from *marbles6*. This means the best that any model can do in predicting *marbles1* is 27%, whereas in *marbles6* it is only 10%. Therefore, it would be misleading to compare prediction rates between data sets that differ in variability because models fit to variable data will always be penalized.

I have not come across any existing literature that addresses this problem (again, because model fits are usually compared within one data set and this problem does not arise) so I have developed an adjusted prediction rate, which

should be used as a guide in comparing model fits across different data sets. The adjusted prediction rate is calculated by dividing the prediction rate by its ceiling:

$$Adjusted\ prediction\ rate = \frac{prediction\ rate}{prediction\ rate\ ceiling} \tag{5.11}$$

In all model fit results, I will report the adjusted prediction rate in parentheses next to the real prediction rate.

5.4 Model fitting results for Experiment 1

5.4.1 Biases underlying probability matching behavior

In this section, we look at how the models fit the probability matching behavior obtained in Experiment 1. Figure 5.6 shows the log likelihood calculations for the range of priors explored ($0 < \frac{\alpha}{2} \leq 100$) for the sampler, averager, and maximizer models.[5] The maximum likelihood estimate of $\frac{\alpha}{2}$ is the value at the peak of each curve and corresponds to the prior that best explains participants' productions.

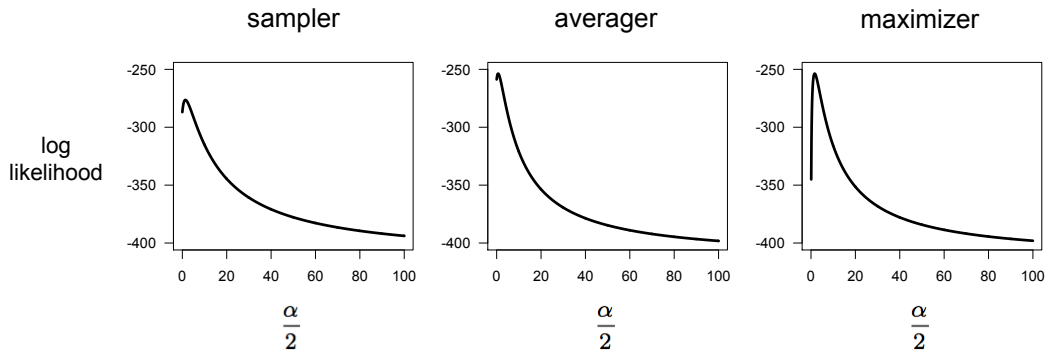


Figure 5.6: Log likelihoods of Experiment 1 data for different values of $\frac{\alpha}{2}$ under the sampler, averager, and maximizer models.

	best-fit $\frac{\alpha}{2}$			log likelihood			prediction rate		
model	sam.	avg.	max.	sam.	avg.	max.	sam.	avg.	max.
best-fit prior	1.39	**0.66**	**1.66**	-276	**-254**	**-254**	17% (49%)	**20% (57%)**	**20% (57%)**

Table 5.1: Maximum likelihood estimates of the best-fit prior parameter $\frac{\alpha}{2}$ for participants in Experiment 1.

The best fit values of $\frac{\alpha}{2}$ are reported in Table 5.1 along with their corresponding log likelihood and prediction rate. The model that returns the best fit overall is indicated in bold. Both the sampler and maximizer models yield a best-fit prior with a weak variability bias ($\frac{\alpha}{2} = 1.39$ and 1.66, respectively) and the averager model yields a prior with a weak regularity bias ($\frac{\alpha}{2} = 0.66$). All of the models do fairly well at explaining the data, but the averager and maximizer models do best with a 20% prediction rate.

[5]The reason I have explored such a wide space of variability-biased priors is because the space of priors itself is skewed. The prior favoring maximum-variability lies at $\frac{\alpha}{2} = \infty$, whereas the prior favoring maximum-regularity is $\frac{\alpha}{2} = 0$ (only one unit away form the unbiased prior $\frac{\alpha}{2} = 1$). Priors favoring maximum-variability are all binomials where $p = 0.5$ and certainly lie within the scope of participant behavior. This prior would be returned by the model fit if, for example, participants were responding randomly in the production phase. Because this is the prior that corresponds to "nothing matters, this experiment is boring", I will include large values of $\frac{\alpha}{2}$ in the search range as a precaution.

observation ratio	5:5	6:4	7:3	8:2	9:1
mode of responses	0.5	0.6	0.7	0.8	0.9
mean of responses	0.47	0.56	0.67	0.77	0.86

Table 5.2: The mode and means of participant responses per observation ratio in Experiment 1. The modes equal the observation proportion of *variant* x (from observation ratio $x : y$) and the means are all slightly biased in the direction of the 5:5 production ratio.

Intuitively, we would expect probability matching to be best described by an unbiased prior, $\frac{\alpha}{2} = 1$, however none of the models return an unbiased prior as the best fit. Although the data in Experiment 1 do qualify as unbiased probability matching behavior because the mode of participant responses is on the input ratio and there was no significant difference between the mean and the input ratio (refer back to the definition on page 48), the raw data was slightly skewed toward the 5:5 ratio (see Table 5.2). This may be what the sampler and maximizer models are picking up on when returning a weak variability bias as the best-fit prior. However in the case of the averager, a weak regularity bias seems largely unfounded on the basis of this data. (Over the course of this chapter we will see that the averager provides the least intuitive results, and the reasons behind this will be discussed further in Section 5.9). The sampler and maximizer fit indicates that participants possess a weak variability bias when reasoning about the ratio of marbles in containers and this is supported by the observation that humans tend to expect random events to be uniformly distributed (Kahneman and Tversky, 1972). A bias toward a 5:5 ratio is evidence of an expectation that two marble colors are uniformly distributed. As anecdotal support for this claim, one participant who received the 8:2 observation ratio left a comment saying "I really do hope the marbles were random. It was crazy how many blues there were!" Incidentally, this person also produced a 8:2 ratio. If Experiment 1 were run with a larger sample size, mean responses may turn out to be significantly lower than the input ratios and support a variability bias interpretation. However, we can not draw that conclusion given the present data set.

Comparing these model fitting results to the coin flipping experiment of Reali and Griffiths (2009), their best-fit prior for the sampler model was $\frac{\alpha}{2} = 4.38$, which is a stronger variability bias than those returned for the marble drawing data. I chose to use stimuli of marbles and containers because people are likely to have a different prior over the possible proportions of marbles in bags than the possible weightings of heads and tails on a coin. Participants are likely to have had experience with variable proportions of marbles in bags, but in the case of coin flipping, especially with real coins as used in Reali and Griffiths (2009), par-

ticipants are likely to only have experience with evenly weighted coins and thus, possess a strong prior bias toward $\theta = 0.5$. They showed that this was the case; iterated learning chains of participants which were initialized with a 9:1 observation ratio, quickly converged to 5:5 production ratios, despite strong evidence in the initial data that the coin was biased. From personal communication with one of the authors, when this coin flipping task was piloted on a computer, where heads and tails images appeared on a the screen, participants probability matched more. This makes sense because participants may have less reason to believe that a computer-generated series of coin flips is constrained by the physical nature of coins and thus, the computer-based framing of the task may have prompted participants to entertain hypotheses differently. Experiment 1 revealed a different kind of prior, one closer to probability matching, because participants' priors within the marble drawing task framing were not strong enough to overwhelm the data they had observed.

5.5 Model fitting results for Experiment 2

Now, we will look at the experimental manipulations that take participant behavior away from probability matching and see what these models tell us about the biases underlying participants' regularization behavior in Experiment 2. This experiment consisted of four conditions: a replication of Experiment 1 called *marbles1*, a replication of Experiment 1 with linguistic stimuli (synonyms and objects instead of marbles and containers) called *words1*, and two multiple frequency learning conditions, in which participants learn about six containers or six objects concurrently, called *marbles6* and *words6* respectively. The two manipulations in this 2×2 design are domain and concurrency and both elicited regularization behavior.

5.5.1 Domain-specific regularization biases

In this section, the sampler, averager, and maximizer models are fit to participant data pooled by domain (marbles vs. words) to determine whether participants possess different prior expectations about the relative frequencies of marbles in containers versus synonyms for objects. In short, does the difference in task framing by domain trigger different frequency learning biases? The best-fit values for prior parameter $\frac{\alpha}{2}$ are shown in Table 5.3.

	best-fit $\frac{\alpha}{2}$			log likelihood			prediction rate		
data set	sam.	avg.	max.	sam.	avg.	max.	sam.	avg.	max.
marbles1 & *marbles6*	**0.43**	1.55	2.55	**-1136**	-1359	-1359	**9% (41%)**	6% (27%)	6% (27%)
words1 & *words6*	$\approx$ **0**	≈ 0	0.92	**-1170**	-1537	-1537	**9% (24%)**	4% (11%)	4% (11%)

Table 5.3: Maximum likelihood estimates of the best-fit prior parameter $\frac{\alpha}{2}$ per domain for participants in Experiment 2.

Each model fits a more regular bias to the words domain than the marbles domain, indicating that there is less of a regularization bias underlying the production of these data compared to those of the words domain, and this is consistent with participant behavior across domain. In the words domain, all of the best-fit priors are regularization biases. For the sampler and averager models, the best-fit bias is the maximally-regular bias of $\frac{\alpha}{2} \approx 0$, and for the maximizer model, a weak regularization bias is the best fit. This is consistent with recent findings that weak inductive biases are sufficient for maximizers to produce strongly biased behavior (Griffiths and Kalish, 2007; Kirby et al., 2007; Smith and Kirby, 2008; Thompson et al., 2012). On the other hand, the results of the sampler and averager models suggest that there are strong regularization biases for the linguistic domain. This interpretation is in line with the argumentation of Bickerton (1984), although a Bayesian prior need not be interpreted as an innate bias. The prior represents *all* of the knowledge that the learner brings to the task, before any data for the task at hand is seen (Griffiths and Kalish, 2007). Therefore, the prior could represent a regularization bias acquired during language acquisition, or even something as short-term as a priming effect.

In the marbles domain, the sampler's bias is weakly regular, while the averager and maximizer models both show variability biases. This could be because the later two models are picking up on the tendency of participants to bias their marble drawing ratios slightly toward the 5:5 ratio, as found in Experiment 1, whereas the sampler is more sensitive to the regularity obtained in *marbles6*.

5.5.2 Demand-based regularization biases

In this section, the models are fit to participant data pooled by the concurrency manipulation (single vs concurrent frequency learning) to determine the prior bias associated with these conditions. It is a little less intuitive to ask about differences in prior expectations due to the concurrency manipulation than it was for the domain manipulation. Because this manipulation increases cognitive load for participants, regularization behavior here is likely due to memory constraints. Although these constraints can impose a bias during induction, the Bayesian model carves up induction with an explicit bias on frequency estimation only, and this may not be the same, or only source of regularization bias produced by memory constraints. Whatever is going on to elicit different regularization behavior between these two conditions, the Bayesian model will respond to participant regularization behavior by capturing it in the prior parameter, as if it were a prior bias on the estimates of frequencies. So the model fits presented in this section should be understood as a catch-all indicator of participants' inductive biases due to cognitive load manipulation.

Table 5.4 gives the model fitting results to the Experiment 2 data pooled by single frequency learning (*marbles1* & *words1*) and concurrent frequency learning (*marbles6* & *words6*).

	best-fit $\frac{\alpha}{2}$			log likelihood			prediction rate		
data set	sam.	avg.	max.	sam.	avg.	max.	sam.	avg.	max.
marbles1 & *words1*	**0.47**	0.66	1.66	**-657**	-720	-720	**13% (33%)**	11% (28%)	11% (28%)
marbles6 & *words6*	$\approx$ **0**	0.59	1.59	**-1650**	-2210	-2210	**8% (33%)**	3% (13%)	3% (13%)

Table 5.4: Maximum likelihood estimates of the best-fit prior parameter $\frac{\alpha}{2}$ per domain for participants in Experiment 2.

The fits indicate that participants have stronger regularization biases for the concurrent frequency learning conditions than the single frequency learning conditions. The sampler model fits a weak regularization bias to single frequency learning and a fully regular bias to concurrent frequency learning. The averager fits a weak regularization bias to both conditions, with concurrent frequency learning biased more toward regularity. The maximizer fits a weak variability bias to both, with concurrent frequency learning less biased toward variability than single frequency learning. In all cases, the sampler is the better fit again, with a prediction rate of 13% for single frequency learning and 8% for concurrent frequency learning.

5.5.3 Best-fit biases per condition

In this section I present the model fits to each of the four conditions separately to see how the models capture the graded increase in regularization behavior across these conditions. Table 5.5 shows these model fitting results per condition.

	best-fit $\frac{\alpha}{2}$			log likelihood			prediction rate		
data set	sam.	avg.	max.	sam.	avg.	max.	sam.	avg.	max.
marbles1	**1.74**	1.66	2.66	**-314**	-324	-324	**14% (47%)**	13% (33%)	13% (33%)
marbles6	$\approx$ **0**	1.5	2.5	**-811**	-1034	-1034	**8% (44%)**	4% (22%)	4% (22%)
words1	$\approx$ **0**	≈ 0	0.96	**-332**	-384	-384	**13% (26%)**	9% (18%)	9% (18%)
words6	$\approx$ **0**	≈ 0	0.9	**-838**	-1152	-1152	**7% (23%)**	3% (10%)	3% (10%)

Table 5.5: Best-fit prior per condition in Experiment 2.

The first row shows the fits for *marbles1*. This was the only condition in this experiment where participants probability matched and all models fit this data with a variability bias and predict the data equally well, at 14% and 13% percent. The next two conditions, *marbles6* and *words1*, elicited similar amounts of regularization behavior, but yield different model fits for the averager and maximizer. This may be because the mass on production ratios are distributed differently between conditions and correspond to different likelihoods under each model. In *words1*, mass is mostly on fully-regular responses and perfect probability matching, whereas in *marbles6* it is mostly on fully-regular responses with large amounts of noise in between. The models distribute likelihood across behaviors differently, and the sampler seems to be responding more to the fully-regular responses (by assigning the same, maximally regular bias to these two conditions), the maximizer is responding to the probability matching component of *words1* and the noise component of *marbles6*, and the averager's response is somewhere in between. *Words6* is best fit by a regularization bias for all models, but again the maximizer only assigns a weak bias. Considering the two word learning conditions together, there is very little difference in the model fits. This indicates that the inductive bias toward regularity is fixed in the words domain and corroborates the findings in Chapter 3 that language-specific biases are across the board and do not interact with the domain-general biases elicited by concurrent frequency learning.

The sampler model is best at predicting the data in all four conditions, however it shows little sensitivity to the different levels of regularity in these data sets, fitting all conditions in which participants regularized some amount with the maximally regular prior $\frac{\alpha}{2} = 0$. The maximizer model does the best at responding

to the different levels of regularization behavior per condition and returns best-fit inductive biases that match the increasing level of regularity obtained across these four conditions (refer back to Figure 3.13)

5.6 Model fitting results for Experiment 3

This experiment is a replication of *marbles6* for two additional sets of observation ratios. It consisted of two conditions: one in which all observation ratios were 10:0, and one in which all were 5:5. Because I am excluding 10:0 data to obtain fits for maximizers, I will just present the results of the *all 5:5* condition here. Because the results of the *all 5:5* were not significantly different from those of the 5:5 ratio in *marbles6*, I would expect the models to return the same best-fit priors as they for the whole *marbles6* data set. However, this was not the case.

	best-fit $\frac{\alpha}{2}$			log likelihood			prediction rate		
model	sam.	avg.	max.	sam.	avg.	max.	sam.	avg.	max.
all 5:5	$\approx$ **0**	≈ 0	≈ 0	**-501**	-610	-610	**7% (41%)**	4% (24%)	4% (24%)
marbles6	$\approx$ **0**	1.5	2.5	**-811**	-1034	-1034	**8% (44%)**	4% (22%)	4% (22%)

Table 5.6: Best-fit priors for Experiment 3 (top) and the fits for *marbles6* in Experiment 2 reprinted for comparison.

All models respond with a maximally-regular best-fit prior. Of all the observation ratios, the probability of fully-regular responses is lowest given a 5:5 observation ratio for all models. Because the maximally regular prior assigns the most likelihood to these rare events (which are actually quite frequent in the *all 5:5* data set), it is the best-fit prior for all models. These model fits seem inadequate because participants could have produced many more regular responses than they did in the *all 5:5* condition, and this shows that the models are not sensitive to the diverse levels of regularization that are possible here. The sampler model provides the best fit here because it assigns more likelihood to the fully regular production ratios than the averager and maximizers do and thus, it scores a higher likelihood and prediction rate for this data set.

Due to the limited diversity of this data set, the Bayesian model has responded with a best-fit prior for the 5:5 region of the model's behavior only. This touches on an import point in both model fitting and experimental design, which is that a diversity of observation data is required to make better estimates of participants' inductive biases. There are two parts to this explanation and I will go over them briefly here as an aside.

First, the models return different priors for individual observation ratios. Table 5.7 shows the model fit results for each observation ratio in *marbles6*.

	best-fit $\frac{\alpha}{2}$			log likelihood			prediction rate		
model	sam.	avg.	max.	sam.	avg.	max.	sam.	avg.	max.
5:5	≈ 0	≈ 0	≈ 0	-171	-211	-211	7%	4%	4%
6:4	≈ 0	2.11	3.11	-183	-238	-238	6%	2%	2%
7:3	≈ 0	≈ 0	0.27	-141	-165	-163	11%	8%	8%
8:2	≈ 0	0.89	1.89	-149	-191	-191	10%	5%	5%
9:1	0.78	2.49	3.49	-163	-218	-218	8%	3%	3%

Table 5.7: Best-fit priors for each observation ratio in *marbles6*.

All of the models return a best-fit prior of ≈ 0 for the 5:5 observation ratio data. This tells us that the models are not treating the *all 5:5* any differently than they are the *marbles6* data. So perhaps, participants' inductive biases do operate similarly in these two experiments.

Second, the global fit to all of the individual observation ratios is not a simple average of the individual best-fit priors. The average of the priors in Table 5.7 is 0.156 for the sampler, 1.098 for the averager, and 1.752 for the maximizer. In Section 5.3 I discussed how the log likelihood scores were constrained by the entropy of the each data distribution. These likelihood scores are also constrained by the entropy of the model's distribution of behavior, and this varies depending on the observation ratio. The balance point between these two ceilings (how much likelihood can be assigned to the data and can be assigned by the model) determines the weighting on the averaged priors.

5.7 Model fitting results for Experiment 4

This is perhaps the most interesting data set to fit the Bayesian models to. Experiment 4 was a replication of *words6* where cognitive load was lightened for participants by either blocking observation trials or blocking production trials. Although this manipulation did not lead to a significant difference in regularization behavior, according to the linear mixed effects regression analysis, it did yield a significant difference in participants' estimates of what they thought the generating ratio was for the data they observed. Participants in the *observation in blocks* condition reported estimates consisting more of the variable ratios (toward the 5:5 end of the ratio set) than participants in the *production in blocks* condition. In asking participants directly about their estimates, I am asking them about what their prior is. Although these self-reports will not accurately reflect

the prior, they provide another source of information on what it might be. So in this section, I will compare the best-fit priors to participants' self-reported priors.

	best-fit $\frac{\alpha}{2}$			log likelihood			prediction rate		
model	sam.	avg.	max.	sam.	avg.	max.	sam.	avg.	max.
observation in blocks	$\approx \mathbf{0}$	1.49	2.49	**-1002**	-1443	-1443	**4% (18%)**	1% (5%)	1% (5%)
production in blocks	$\approx \mathbf{0}$	0.19	1.19	**-871**	-1234	-1234	**7% (30%)**	2% (9%)	2% (9%)

Table 5.8: Best-fit priors for Experiment 4.

The sampler model returns a prior favoring maximum regularity for both conditions (due to its lack of sensitivity for the range of possible regularization behavior, as mentioned in the previous section) and therefore can not tell us much about the difference in participants' self-reported priors. However, the averager and maximizer models both return priors that are in line with participants' ratio estimates in these conditions: they return a prior that favors more variation in the *observation in blocks* condition than in the *production in blocks* condition. Participants' ratio estimates in *observation in blocks* were clearly biased toward variability, and in *production in blocks* they were clearly unbiased (refer back to Figure 3.22). The averager and maximizer return a variability-biased prior in *observation in blocks*, however neither capture the unbiased estimates in *production in blocks*: the averager model fits these with a weakly regular prior and the maximizer fits these with a weakly variable prior. Although the absolute bias strengths do not equal the actual bias in the self-reported estimates, with-in model rankings are consistent with them.

Because Bayesian model fits rely entirely on participant behavior, this model must be picking up on the nuanced differences in regularization behavior to recover these biases in participants' estimates, despite the inability of the linear mixed effects regression analysis to detect a significant difference in regularization behavior across these two conditions. However, the model fits tell us that the experimental manipulation has affected the data, and done so in a way that is consistent with a Bayesian model's instantiation of participants' inductive bias on frequency estimation.

5.8 Model fitting results for Experiment 5

5.8.1 Biases underlying coordination behavior

In Experiment 5, participants worked with a partner to coordinate on a marble color on the basis of shared experience with draws from a bag, which provided common ground color and relative frequency information. Data were collected for two observation ratios only, 7:3 and 5:5, in separate conditions. Table 5.9 gives the model fitting results for these data pooled together (top row), just the 5:5 observation ratio condition (middle row), and just the 7:3 observation ratio condition (bottom row).

	best-fit $\frac{\alpha}{2}$			log likelihood			prediction rate		
model	sam.	avg.	max.	sam.	avg.	max.	sam.	avg.	max.
all data	**0.18**	1.98	2.98	**-139**	-154	-154	**10% (67%)**	8% (53%)	8% (53%)
5:5	$\approx$ **0**	$\approx$ 0	$\approx$ 0	**-67**	-72	-72	**11% (73%)**	9% (60%)	9% (60%)
7:3	**0.32**	1.98	2.98	**-72**	-83	-83	**9% (60%)**	6% (40%)	6% (40%)

Table 5.9: Best-fit priors for Experiment 5.

The 5:5 condition is best-fit by a fully regular bias by all models (refer back to Section 5.6 for comments the models' lack of sensitivity to regularization behavior from 5:5 observation ratios). For the 7:3 condition, the sampler returns a prior weakly biased for regularity and the averager and maximizer models return a prior weakly biased for variability. The variability-biased priors result from the high number of 5:5 ratios participants produce in this task. The analyses of Experiment 5 showed that this bias toward 5:5 ratios was the result of a bias toward regularity at the level of whole sequences (i.e. the low conditional entropy of perfectly alternating sequential color choices). However, this shows up simply as a bias for 5:5 ratios to the Bayesian model. Model fits to the pooled data (top row) follow the 7:3 pattern. The averager and maximizer model fits are completely driven by the more regular observation ratio (see Section 5.3 for an explanation of this) and the sampler is driven by both observation ratios.

This experiment explicitly manipulated the production phase of the basic frequency learning task used throughout this thesis to see how higher-cognitive decision strategies lead to regularization behavior. Because this Bayesian model defines a prior on frequency estimates, it can only respond to participant data as if it were the result of a bias on the *estimation* procedure. However, information about the inductive biases at play in coordinating humans, whether they be estimation biases or higher-level inferences about what type of data should be

produced in this coordination task, will be reflected (in one way or another) in the Bayesian model fit. When the mapping between the architecture of human inductive biases and the way the model carves up the inference process does not correspond, this may lead to best-fit priors that are difficult to interpret. This issue will be addressed further in the discussion section.

5.9 A closer look at the model's behavior

Now that we have seen the model fit results for each experiment, let's take an in-depth look at the range of behaviors that the sampler, averager, and maximizer models produce. Elaborate descriptions of the model's behavior are not found in the existing literature on this beta-binomial model of frequency estimation, so the facts presented in this section are my own observations and descriptions, and are customized to the parameters I've used in my experiments and the measure of regularization I have employed. In this section, I will examine the relationship between *biased priors* and the *biased behaviors* that result from them and make frequent use of these two distinct concepts. Further ahead, in Figure 5.10, I will plot the relationship between different values of $\frac{\alpha}{2}$ and the exact level of regularization behavior they give rise to. In the discussion section, I will assess each model's ability to describe participant behavior in light of the model fit results and the types of behavior that each of the models are capable of producing.

Figures 5.7, 5.8, and 5.9 show the behavior of the sampler, averager, and maximizer models, respectively, for three example values of $\frac{\alpha}{2}$. Each row corresponds to one value of $\frac{\alpha}{2}$: 0.01, 1, and 100. Each column corresponds to one of the six observation ratios, ranging from 5:5 (on the left) to 10:0 (on the right). The model's *behavior* is defined by the probabilities with which the model produces each possible ratio, given each possible observation ratio (just as participant behavior was described by a probability distribution over production ratios). One of the first things to notice is that all of these models behave similarly when $\frac{\alpha}{2}$ is high (the bottom row of each model, where $\frac{\alpha}{2} = 100$, is nearly identical). In fact, when $\frac{\alpha}{2} = \infty$ model behavior *is* identical: all distributions are binomials where $p = 0.5$, regardless of the observation ratio. Table 5.10 prints the mean of each distribution and Table 5.11 prints the variances. On all three models for $\frac{\alpha}{2} = 100$ (where 100 is a good proxy for ∞) the mean and variance of these distributions conform to the binomial mean, $pN = 0.5$, and variance, $p(1-p)N = 2.5$.

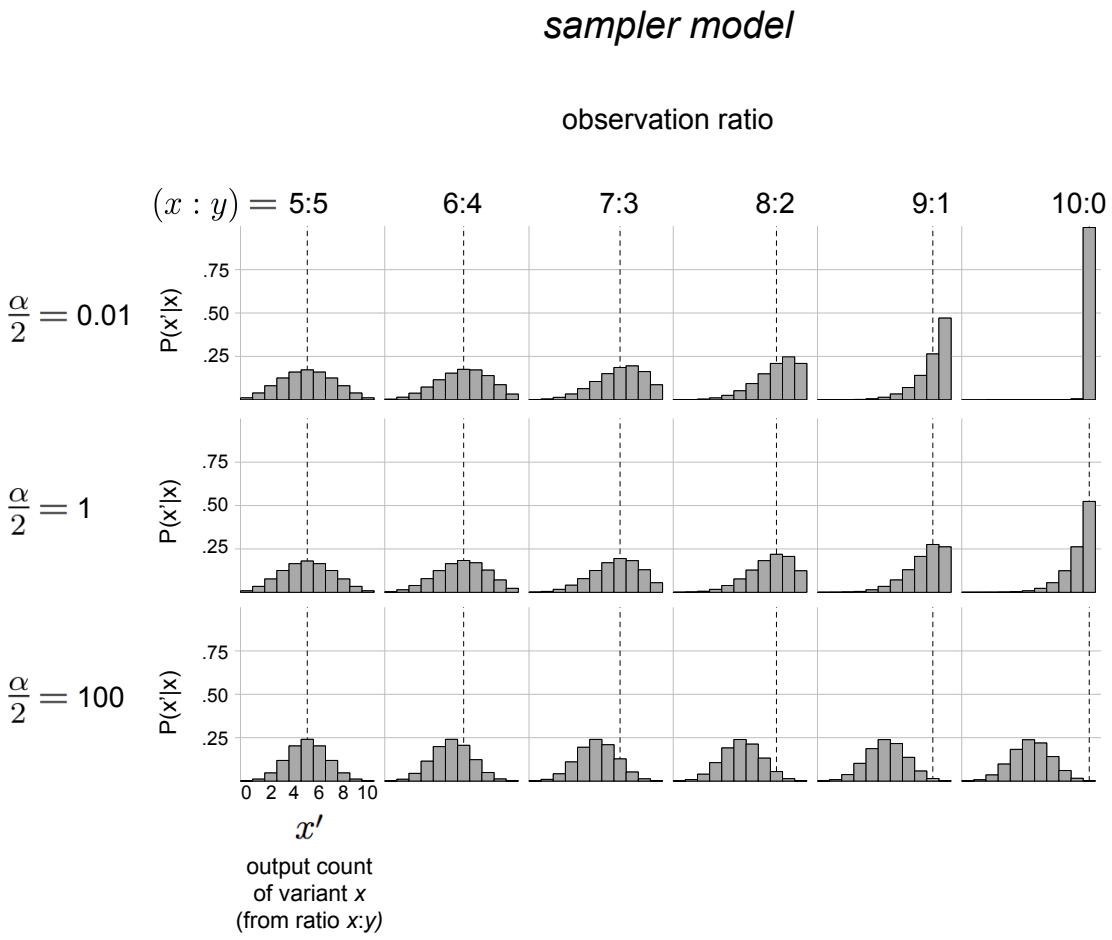


Figure 5.7: Behavior of the sampler model, for three example values of $\frac{\alpha}{2}$: 0.01 (regularization bias), 1 (unbiased), 100 (variabilization bias). Each row corresponds to one value of $\frac{\alpha}{2}$. Each column corresponds to one of the six observation ratios, ranging from 5:5 (on the left) to 10:0 (on the right). Each panel contains the distribution of ratios that the model would produce in response to one observation ratio. These production ratios are displayed on the x-axis as the number of productions of *variant x* from the observation ratio *x:y*. *Variant x* corresponds to whatever variant was in the majority during the observation phase. All observation ratios are indicated by a dashed line. For example, the top left panel shows the behavior where $\frac{\alpha}{2} = 0.01$ in terms of a probability distribution over all possible productions. Here, the probability of producing *variant x* 5 times is 0.17. There's a 0.16 probability of producing it 6 times, and a 0.13 probability of producing it 7 times. Because a symmetrical prior is used in these models, the observation ratios not shown (4:6, 3:7, 2:8, 1:9, 0:10) are all mirror images of 6:4, 7:3, 8:2, 9:1, 10:0, respectively.

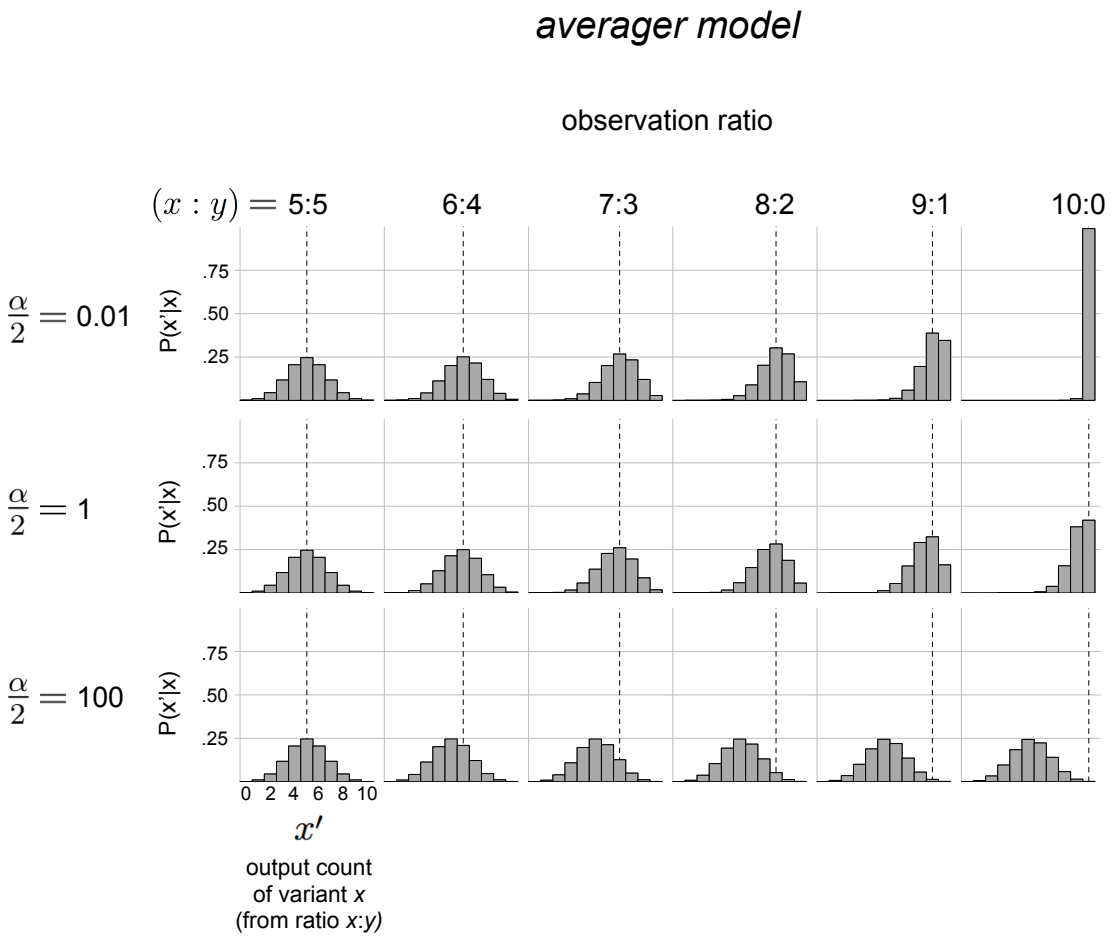


Figure 5.8: Behavior of the averager model, for three example values of $\frac{\alpha}{2}$: 0.01 (regularization bias), 1 (unbiased), 100 (variabilization bias). Each row corresponds to one value of $\frac{\alpha}{2}$. Each column corresponds to one of the six observation ratios, ranging from 5:5 (on the left) to 10:0 (on the right). Each panel contains the distribution of ratios that the model would produce in response to one observation ratio. These production ratios are displayed on the x-axis as the number of productions of *variant x* from the observation ratio *x:y*. *Variant x* corresponds to whatever variant was in the majority during the observation phase. All observation ratios are indicated by a dashed line.

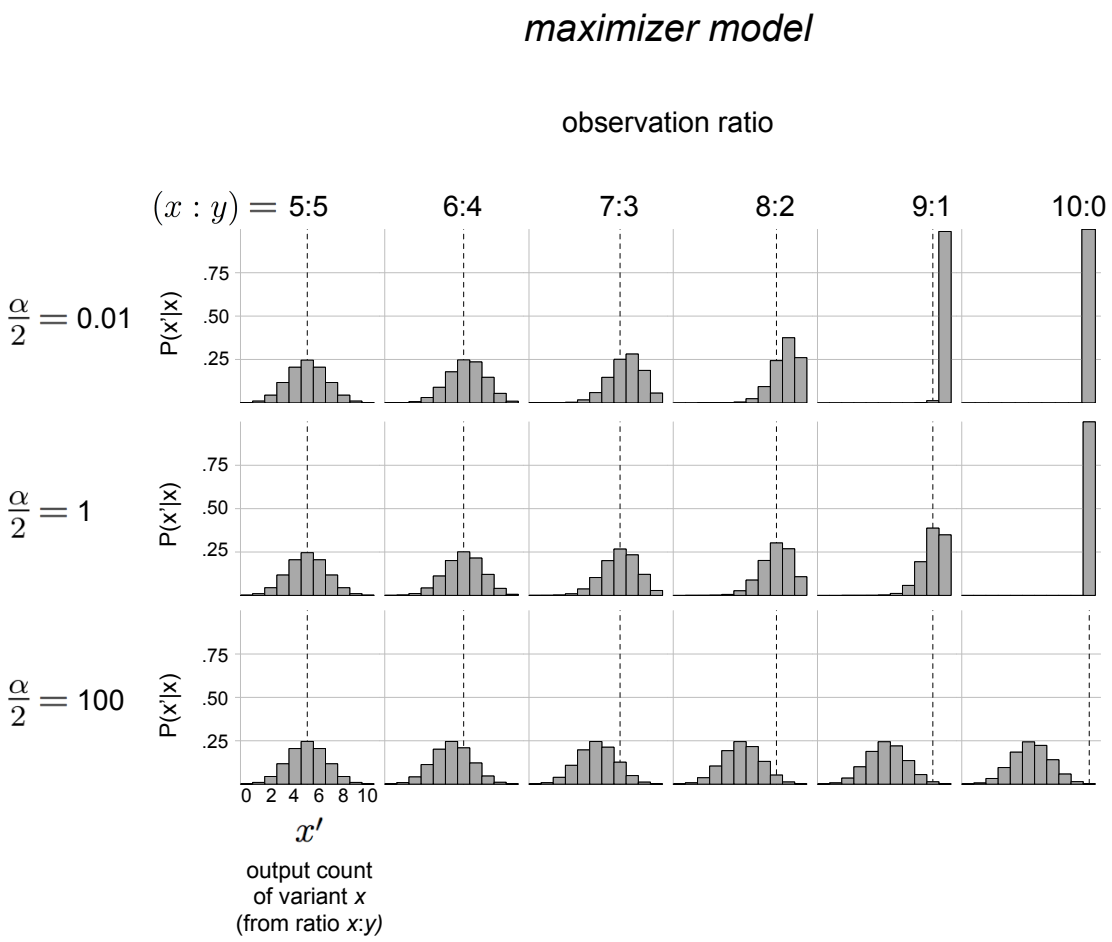


Figure 5.9: Behavior of the maximizer model, for three example values of $\frac{\alpha}{2}$: 0.01 (regularization bias), 1 (unbiased), 100 (variabilization bias). Each row corresponds to one value of $\frac{\alpha}{2}$. Each column corresponds to one of the six observation ratios, ranging from 5:5 (on the left) to 10:0 (on the right). Each panel contains the distribution of ratios that the model would produce in response to one observation ratio. These production ratios are displayed on the x-axis as the number of productions of *variant x* from the observation ratio *x:y*. *Variant x* corresponds to whatever variant was in the majority during the observation phase. All observation ratios are indicated by a dashed line. The distribution for a 10:0 observation ratio when $\frac{\alpha}{2} > 1$ is undefined, because the mean is greater than 10. However, it is plotted here with the maximum mean, 10.

When the prior is unbiased ($\frac{\alpha}{2} = 1$), we would expect the model's behavior to be unbiased as well. Unbiased behavior is defined by distributions where the mean equals the observation count of *variant x*. Surprisingly, only the maximizer model produces unbiased behavior when the prior is unbiased. The means of each distribution that the maximizer with $\frac{\alpha}{2} = 1$ produces are all equal their respective observation count of *variant x* (Table 5.10, third to last column). Additionally, the mean and variances of the maximizer with $\frac{\alpha}{2} = 1$ show that this model is equivalent to binomial drift, because they are identical to the mean (pN) and variance ($p(1-p)N$) of binomial drift (see the last column of Tables 5.10 and 5.11). In fact, for all values of N, a maximizer with $\frac{\alpha}{2} = 1$ is equivalent to binomial drift (not shown). This is an important point that will be taken up in the discussion, in relation to the Wright-Fisher drift equivalence results of Reali and Griffiths (2010).

As for the sampler and averager models, when they have unbiased priors their behavior is *always* biased toward variable responses. This is because both of their hypothesis selection strategies choose the average of the posterior distribution (the averager chooses it deterministically and the sampler chooses makes stochastic hypothesis choices that are evenly distributed about the posterior mean) and this average corresponds to a more variable value of θ than the mode (except for when the average equals the mode at $\frac{\alpha}{2} = 0$). Thus, only when $\frac{\alpha}{2} = 0$ do these two models produce unbiased behavior. This is problematic given our understanding of regularization and shows that these two models implement a counter-intuitive relationship between the prior bias and behavioral bias. Also, this is why most of the sampler and averager best-fit priors were $\frac{\alpha}{2} = 0$ for the data sets where participants regularized: these models can not capture regularization in the sense of a behavioral bias toward over-producing the majority variant.[6]

The maximizer, on the other hand, is capable of mean overproduction of the majority variant. When $\frac{\alpha}{2} < 1$, means are all greater than the observed count of *variant x*. This is why the maximizer model was able to fit the regularization data with various strengths of a regularity-biased prior: this model has a wider range of regularization behavior that allows for finer sensitivity to the different levels of regularity produced by participants in the different experimental conditions.

In Table 5.10 it can be seen that the sampler and averager models always behave with identical means.[7] This is because the hypothesis selection strategy of

[6] As far as I am aware, this has not been discussed on in the existing literature on this model, so I am not sure what to make of this, besides just throwing out these models and sticking to the maximizer for describing regularization behavior.

[7] This is the analytical solution for an infinitely large population.

	sampler			averager			maximizer			binomial
$\frac{\alpha}{2} =$	0.01	1	100	0.01	1	100	0.01	1	100	
5:5	5.000	5.0	5.0	5.000	5.0	5.00	5.0	5.0	5.00	5.0
6:4	5.998	5.8	5.05	5.998	5.8	5.05	6.2	6.0	5.05	6.0
7:3	6.996	6.7	5.10	6.996	6.7	5.10	7.5	7.0	5.10	7.0
8:2	7.994	7.5	5.14	7.994	7.5	5.14	8.7	8.0	5.14	8.0
9:1	8.992	8.3	5.19	8.992	8.3	5.19	10.0	9.0	5.19	9.0
10:0	9.990	9.2	5.24	9.990	9.2	5.24	11.2	10.0	5.24	10.0

Table 5.10: The mean of each distribution in Figure 5.7, Figure 5.8, Figure 5.9. The rightmost column gives the mean of the binomial distribution where $p = \frac{x}{N}$ of each observation ratio *x:y*.

	sampler			averager			maximizer			binomial
$\frac{\alpha}{2} =$	0.01	1	100	0.01	1	100	0.01	1	100	
5:5	4.54	4.23	2.61	2.50	2.50	2.50	2.5	2.50	2.50	2.50
6:4	4.36	4.11	2.61	2.40	2.43	2.50	2.34	2.40	2.50	2.40
7:3	3.82	3.76	2.61	2.10	2.22	2.50	1.88	2.10	2.50	2.10
8:2	2.91	3.17	2.60	1.60	1.88	2.50	1.10	1.60	2.50	1.60
9:1	1.65	2.35	2.60	0.91	1.39	2.50	0.01	0.90	2.50	0.90
10:0	0.02	1.29	2.60	0.01	0.76	2.50	undef.	0.00	2.50	0.00

Table 5.11: The variance of each distribution in Figure 5.7, Figure 5.8, Figure 5.9. The rightmost column gives the variance of the binomial distribution where $p = \frac{x}{N}$ of each observation ratio *x:y*.

the sampler is symmetric about the posterior mean, producing binomial distributions of data from a hypothesis sampled evenly about the mean. The averager's hypothesis selection strategy deterministically chooses the posterior mean and produces data from this mean via binomial sampling. Only the maximizer model differs in respect to mean behavior per observation ratio, because its hypothesis selection strategy takes the mode of the posterior. Therefore, the mean of the maximizer's behavior is offset from the mean sampler and averager behavior by the difference between the mode and the mean of the posterior distribution:

$$\frac{\frac{\alpha}{2} + x - 1}{\alpha + N - 2} - \frac{\frac{\alpha}{2} + x}{\alpha + N}. \tag{5.12}$$

When $\frac{\alpha}{2} = 0$ for the averger, the maximizer produces identical behavior when its $\frac{\alpha}{2} = 1$. Likewise, when the averager's $\frac{\alpha}{2} = 1, 1.5, 2$, or 2.5, the maximizer produces identical behavior at $\frac{\alpha}{2} = 2, 2.5, 3$, or 3.5, respectively. For all values of $\frac{\alpha}{2}$ on the averager model, the maximizer produces identical behavior with a prior that is greater by one unit. That is why all of the model fit results show a unit difference relationship between the best-fit priors of the averager and maximizer whenever the maxmizer's best-fit prior is $\frac{\alpha}{2} = 1$ or greater (for an example, refer back to the fit results for Experiment 2, Table 5.5, top two rows). When the maxmizer's

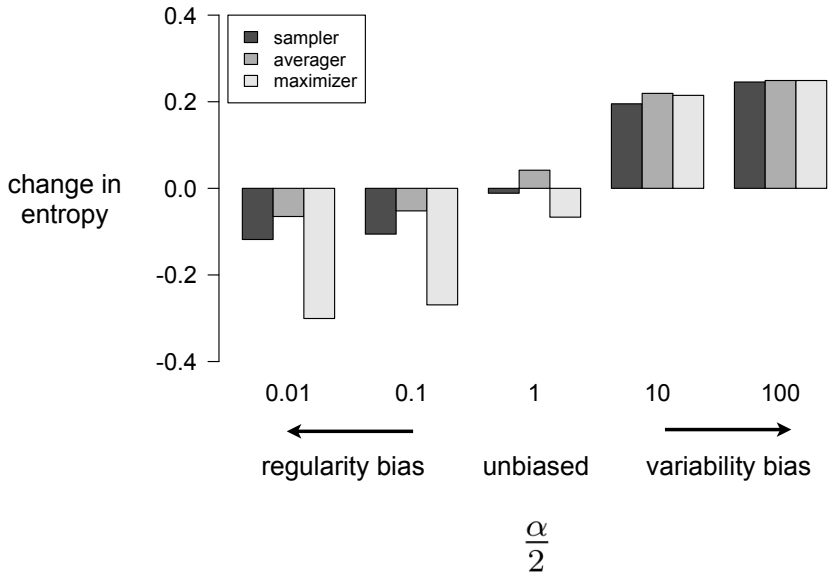


Figure 5.10: Average change in entropy (in bits) for all observation ratio to production ratio transitions in the sampler, averager, and maximizer models for five example values of $\frac{\alpha}{2}$. The models vary in their ability to regularize.

best-fit prior is less than $\frac{\alpha}{2} = 1$, this describes behavior that the averager model is not capable of, and the averager will return a best-fit prior of $\frac{\alpha}{2} = 0$ (again, see Table 5.5, bottom two rows).

In terms of variance (Table 5.11), both the averager and maximizer produce behaviors with binomial variance, because their productions follow binomial distributions. Therefore, these models are equivalent to Wright-Fisher drift for values of $\frac{\alpha}{2}$ that lead to unbiased behavior (for the maximizer, this is when $\frac{\alpha}{2} = 1$ and for the averager, this is when $\frac{\alpha}{2} = 0$). The sampler, however, does not produce binomially distributed data, except when $\frac{\alpha}{2} = \infty$. Besides this one exception, the variance of the sampler is always higher than binomial variance. Therefore, the behaviorally unbiased case of the sampler (when $\frac{\alpha}{2} = 0$)[8] is not equivalent to binomial drift, because these distributions are not binomials, but does constitute a form of cultural drift that is not accounted by the Wright-Fisher model. This is an important point that I will return to in the discussion.

Now that we have seen the behavioral profiles of each model, I will assess the different levels of regularization behavior that these models are capable of. The extent to which a model can regularize is determined by the mass it places on production ratios that have lower entropy than its observation ratio. The models

[8]The quirk in the sampler and averager models, where maximally biased priors lead to unbiased behavior, is correct. Refer back to Table 5.10.

can achieve this in two ways: 1) by biasing the mean of productions toward regular ratios (as the maximizer does) and 2) by increasing the variance and/or skew of productions, such that extremely regular ratios are produced with a higher probability (as the sampler does). For an example of the first point, compare the 8:2 panes for the maximizer where $\frac{\alpha}{2} = 0.01$ and $\frac{\alpha}{2} = 1$ (Figure 5.9). The mean of the distribution for $\frac{\alpha}{2} = 1$ is $x' = 8$, whereas for $\frac{\alpha}{2} = 0.01$ it is higher than x, at $x' = 8.74$ and therefore places more mass on the 10:0 production ratio. For an example of the second point, compare the 8:2 panes for the sampler where $\frac{\alpha}{2} = 0.01$ and $\frac{\alpha}{2} = 1$ (Figure 5.7). The distribution for $\frac{\alpha}{2} = 0.01$, despite having a mean lower than $x' = 8$, is more skewed toward higher values of x' and places more mass on the 10:0 production ratio than the distribution for $\frac{\alpha}{2} = 1$.

Figure 5.10 shows each model's behavior in terms of the entropy change it achieves, for five different values of $\frac{\alpha}{2}$ (0.01, 0.1, 1, 10, 100). Here we see again that all three models perform similarly for high-variability biases, but differ markedly in terms of regularization capabilities. The maximizer is the only model that is able to achieve entropy drops as strong as those achieved by participants. The maximum entropy drop achievable by the maximizer is -0.30 bits when $\frac{\alpha}{2} = 0$, whereas the sampler and averager can only achieve -0.12 and -0.07 bits, respectively. For comparison, the largest entropy drop achieved by participants was in the *marbles6* condition of Experiment 2, at -0.36 bits.

The averager model regularizes little because it neither biases mean productions toward regular ratios, nor places probability on extremely regular ratios via high variance in productions. Compare the three plots in Figure 5.7, 5.8, and 5.9 for the 9:1 observation ratio when $\frac{\alpha}{2} = 0.01$. Compared to the averagers, samplers are much more likely to regularize a 9:1 ratio by producing at 10:0 ratio because their variance is wider, and maximizers are much more likely to produce a 10:0 because their mean is strongly biased toward that ratio. Here, averagers are most likely to produce a 9:1 ratio, leading to no change in entropy and thus, no regularization behavior.

5.10 Discussion

In this chapter, I fit three Bayesian models of frequency learning and production to a diverse set of human frequency learning data in a linguistic and non-linguistic domain, under different levels of cognitive load, and in contexts that modulate participants' frequency production strategies. These data sets came from each of the five experiments presented in this thesis. A variety of cognitive biases operate in individuals as they are learning about the frequencies of events in their environment and each of these experiments was designed to engage a particular subset of these biases. The three main bias types explored with these experiments were biases on participants estimates of frequencies, biases due to memory encoding and recall under different levels of cognitive load, and production biases.

Because many biases are involved in perception, processing, and production, the inductive loop that turns observations into productions can be carved up in many ways. The model applied in this chapter, the beta-binomial Bayesian model of frequency estimation, defines the inductive bias as a bias on frequency estimation. Therefore, the model fits should be more intuitive for the experiments that were more likely to engage frequency estimate biases. However, these models will respond to an inductive bias no matter its source and thus, will pick up on any biases in frequency learning behavior and report it as if it were a bias on frequency estimation. This means that we can still use this Bayesian model as an indicator of participants' inductive biases, but I want to point out that the envelope is being pushed in some cases.

Table 5.12 provides a summary of the main model fitting results presented in this chapter. The first column specifies the data set the model was fit to and the second column gives the predominant bias that each experiment was designed to engage. The third column shows which model (sampler, averager, or maximizer) returned the best fit and the fourth column coarsely codes each model's best-fit prior by whether it was a biased toward regularity (+) or biased toward variability (−). The first thing to notice in this table is that the sampler model is consistently the best fit to all of the data sets that elicit regularization behavior (Experiments 2 through 5). Only in Experiment 1, when participants are probability matching, do the averager and maximizer models do best. This is because the averager and maximizers have a more restricted variance than the sampler model and this allows them to fit the low-variance probability matching behavior of humans better. As for the regularization behavior, why is the sampler the best fit? In the previous section, we saw that the maximizer model was the only model capable of regularizing as much as participants do (refer back to Table

data set		participant bias type	best-fit model	sam.	avg.	max.
Experiment 1		estimate	avg. & max.	−	+	−
Experiment 2	*marbles1*	estimate	sampler	−	−	−
	words1	estimate	sampler	+	−	−
	marbles6	memory	sampler	+	+	+
	words6	memory	sampler	+	+	+
Experiment 3		memory	sampler	+	+	+
Experiment 4						
observation in blocks		estimate & memory	sampler	+	−	−
production in blocks		estimate & memory	sampler	+	+	−
Experiment 5		production	sampler	+	−	−

Table 5.12: Summary table of the main model fitting results from this chapter. The last three columns code the best-fit prior of each model (sampler, averager, maximizer) with a "+" for a regularity bias and a "−" for a variability bias.

5.10) and additionally was the only model capable of regularizing via a bias in mean behavior.

The sampler, on the other hand, is capable of regularizing by assigning more mass to the fully-regular production ratios (10:0 and 0:10) than the averager and maximizer do. Therefore, it provides a better fit to all of the data sets that elicit regularization behavior.[9] It appears that, in general, averagers and maximizers will always provide better fits to probability matching data and samplers will always provide better fits to regularization behavior and probably to all noisy behaviors with fairly uniform distributions over outcomes as well.

As for the types of priors each model returned, the sampler results matched up best with the data: when participants probability matched, the sampler returned a bias toward variability and when participants regularized, it returned a bias toward regularity. Overall, however, the sampler showed little sensitivity to the different amounts of regularization behavior elicited by each experiment and thus, little sensitivity to the different strengths of the biases that may have been engaged. The averager and maximizer show more sensitivity and this paid off in Experiment 4, where they were able to detect a difference in priors that corresponded to participants' self-reported estimates of the marbles in the containers. Participants reported more variable ratios in the *observation in blocks* condition and both models fit a more variable prior to this condition than the

[9]It also provides a better fit to the probability matching behavior in *marbles1*, but only by a hair. The sampler's prediction rate for *marbles1* was 14% whereas the averager's and maximizer's were each 13%.

other (refer back to Table 5.8 for the exact bias strengths). The sampler, on the other hand, returned a maximally strong prior for both of these data sets. Although the averager and maximizer show more sensitivity, they sometimes returned counter-intuitive priors. For example they both fit a variability prior to *words1*, which elicited regularization behavior. This data set was composed of fairly bimodal data in which some participants regularized and others probability matched. Because the averager and maximizer assign likelihood better to probability matching behavior, this probably drove the overall model fit, such that a variability prior won out in this data set. And in Experiment 4 and 5, participants clearly regularized, but the averager and maximizer returned variable priors.

As shown in Section 5.9, none of these models do a particularly good job at capturing truly human-like probability matching and regularization behavior. When the models are probability matching, the peaks on their distributions are too flat. Human participants are remarkably good at reproducing the frequencies they observe, but the Bayesian models perform with much more error. The averager and maximizer models probability match with binomial error and the sampler model probability matches with even more error. The analyses in Chapter 2 showed that human probability matching behavior is significantly better than binomial probability matching. However, the models do much worse at capturing human regularization behavior than they do for probability matching behavior. Humans often regularize by exclusively producing either the majority variant or the minority variant. This yields U-shaped distributions over production ratios with most mass on the fully-regular (10:0 and 0:10) ratios. All of these models are incapable of producing U-shaped distributions over production ratios. Model regularization behavior is unimodal and places most mass near the observation frequency. When the models do place a lot of mass on a fully-regular ratio, it is always that of the majority variant. The poor fit of model to human regularization behavior is exemplified by the 5:5 observation ratios. Human regularization of 5:5 ratios is U-shaped whereas the model's never is. The models only seem to produce human regularization-like behavior for observation ratios that are already fairly regular (i.e. the 8:2, 9:1 ratios) whereas for the variable ratios (i.e. 5:5, 6:4), model behavior always looks more like probability matching. This indicates that there may be something about human regularization behavior that can not be completely accounted for in terms of a bias on frequency estimates alone.

On the basis of this, I would like to suggest some avenues for future research with this particular Bayesian model. There are other places in this model where

inductive biases, in addition to one on frequency estimation, could occur. The replication of cultural variants occurs over an inductive loop that creates a set of productions from a set of observations and this model bridges observations and productions with four components: the prior bias, the likelihood, the hypothesis selection strategy, and the data production algorithm. Three of these four components are hard-coded, but could conceivably contain biases as well. There already exist three commonly-implemented versions of the hypothesis selection strategy: the sampler, averager and maximizer models (e.g. Griffiths and Kalish, 2007). The maximizer model biases learners toward choosing the most probable hypothesis. Kirby et al. (2007); Smith and Kirby (2008); Thompson et al. (2012) show that when learners are trying to communicate or coordinate with one another, such a bias makes sense because it enables learners to choose the same hypothesis (i.e. arrive at the same conclusion about a given set of data) as one another. As discussed and demonstrated empirically in Chapter 4, maximizing on the basis of frequency information enables participants to solve a coordination task.

However, yet another form of maximizing can occur when learners generate data. This relates to representativeness and is already being addressed in the Bayesian modeling literature (e.g Tenenbaum and Griffiths, 2001; Rafferty and Griffiths, 2010). A Bayesian learner who is trying to be representative could bias their productions toward those that are most likely under the chosen hypothesis. This would improve the ability of any student agent to arrive at the same hypothesis as their teacher. This could be achieved by raising the distribution over data production probabilities to a power. If this were done for the averager and maximizer models presented in this chapter, then the variance on the distribution of production ratios would be lower and would better-approximate human probability matching behavior. Likewise, if a Bayesian rational learner were learning from a representative teacher, they would need to know this (in order to be rational) and have a data likelihood function that is identical to the data production probabilities of the agents' its learning from. Thus, the likelihood could be raised to a power to reflect a learner's knowledge that teacher agents are trying to be representative. However, breaking this strict criteria that the data likelihoods equal the veridical data production probabilities would break the rationality assumption of agents, but may better approximate human behavior during frequency learning.

Additionally, these models could be improved by adding a component that captures constraints on memory encoding and recall. An example along this line has been developed by Perfors (2012) which models memory limitations by dis-

torting or changing the quantity of observations in four different ways: dropping data at random, reconstructing data randomly, reconstructing data according to the prior bias, and multiplying the data distribution by an amount of decay. This model shows that memory limitations that do not distort the data in a biased way will not lead to regularization behavior unless the learner possesses a prior bias for regularization.

Finally, these models of the inductive process tell us something very important about cultural evolution: that the dynamics associated with inductive evolution may constitute a much wider class of evolutionary processes than those formulated for the biological evolution of organisms. Reali and Griffiths (2010) provide an equivalence proof between the Wright-Fisher model of genetic drift with mutation and Bayesian learners that, when the number of cultural variants is two, correspond to the beta-binomial averager model presented in this chapter. This equivalence relies on the fact that the averager produces data distributions that follow a binomial distribution, just as the Wright-Fisher model does. This equivalence can clearly be seen in my report of this model's behavior in Figure 5.8 for the case of Wright-Fisher drift without mutation. Here, when $\frac{\alpha}{2}$ tends toward 0, the model's distributions over production ratios tend toward binomial distributions with means on the input frequency and thus, define binomial drift (refer back to Section 1.5.1). Additionally, we can see that this equivalence also holds for the maximizer model, but in a more intuitive way than shown by Reali and Griffiths (2010). Whereas the averager model is equivalent to Wright-Fisher drift without mutation for a maximally-regular prior, the maximizer model is equivalent to Wright-Fisher drift without mutation for the unbiased prior, $\frac{\alpha}{2} = 1$. Because drift is a form of neutral evolution, we should expect it to occur via an unbiased inductive process. Therefore, the maximizer model appears to be a better choice of model to use for the further development of equivalences to Wright-Fisher models, such as that with selection. The sampler model, on the other hand, does not produce binomially-distributed data and therefore can not conform to Wright-Fisher models. This sampler model's dynamics are part of a larger class of evolutionary dynamics and it would be interesting to determine whether or not this model, or other models of evolution by induction have equivalences with any known forms of biological evolution.

Chapter 6

The cultural evolution of regularity

Up until this point, we have been detailing the cultural evolutionary forces that operate on language by an in-depth analysis of a single generation of learners via psychological experimentation. This enabled me to build a detailed description of the cognitive biases at play in frequency learning and how they shape the distribution of behavior in one generation of learning. This chapter adds cultural transmission to the story. What do cognitive biases tell us about the distribution of cultural variants after several generations of learners (the synchronic variation)? How does this distribution change over time (the diachronic variation)? And how much information does the behavior of one generation of learners carry about the behavior after many generations?

These questions are of great concern to those who research the cultural evolution of various human behaviors and are currently being addressed, perhaps most rigorously, within the language evolution literature. Here, the question of how cognitive biases map on to the distribution of linguistic features (within a language, or across the world's languages) is known as *the problem of linkage* (Kirby, 1999):

> **The problem of linkage**. Given a set of observed constraints on cross-linguistic variation, and a corresponding pattern of functional preference, an explanation of this fit will solve the problem: how does the latter give rise to the former? (p. 20)

There are three main avenues of linguistics research that address this problem: nativism (e.g. Chomsky, 1965, 1982; Marantz, 1995; Grimshaw, 1997), functionalism (e.g. Greenberg, 1963; Bybee, 1985; Cutler et al., 1985; Du Bois, 1987; Comrie, 1989; Croft, 2003), and cultural evolution via iterated learning (e.g. Kirby, 2001,

2013; Hurford, 2011; Kalish et al., 2007; Griffiths et al., 2008). Kirby (1999) argues that the nativist approach specifies a mechanism linking universals to acquisition, where constraints on cross-linguistic variation are a direct consequence of innate constraints on the acquisition and mental representation of language, but that this approach does not account for the appearance of design in the goodness of fit between form and function in language. The functionalist approach focuses primarily on this goodness of fit, by stating that universals should fit the pressures imposed by language use, but does not specify the mechanism by which this fit should come about. However, when we recognize that languages evolve culturally and in doing so, adapt to the minds of learners, this provides both the mechanism that links constraints on acquisition to universals and the explanation for the goodness of fit between form and function. Here, the mechanism is the cultural transmission process and the explanation is the evolutionary forces that lead languages to adapt to learners.

Griffiths and Kalish (2007) address the problem of linkage by characterizing the mapping between cognitive biases and typological distribution in an iterated learning modeling framework with Bayesian agents. In this framework, typological distribution is understood as the distribution over linguistic features (or more broadly, cultural variants) after many generations of learners[1]. The use of Bayesian agents makes the cognitive biases of the learners explicit (as the prior distribution over hypotheses) so that this prior can be directly compared to the distribution over cultural variants after several generations of learners. Results are provided for *samplers* and *MAP learners* (refer back to Chapter 5). When learners are samplers, the distribution over variants comes to mirror the prior distribution over hypotheses exactly, such that the prior fully predicts long-term behavior, and vice versa. This result suggests that the cultural transmission mechanism merely enables an inevitable one-to-one correspondence between typological distribution and learning biases. If the agents' prior is interpreted as an innate bias, then this means that the parameters of the nativists' language acquisition device can be directly read off from the typology of the languages of the world (although see Dunn et al. (2011) for a phylolinguistic account of why historical dependencies among the world's languages confound this conclusion). And if the agents' prior is interpreted as a functional preference for certain variants, then the functionalist account of goodness of fit is complete, for all intents and purposes, because the cultural transmission mechanism does not add anything to the account of this fit. However, this convergence to the prior result seems to be a

[1] more precisely, as the number of generations approaches infinity.

special case, and does not hold when different assumptions of this sampler model are relaxed. Griffiths and Kalish (2007) also show that for MAP learners, this relationship is not one-to-one and tends to over-represent variants corresponding to hypotheses that have higher prior probability. Kirby et al. (2007) detail the non-convergence result of the MAP learner further. Additionally, Ferdinand and Zuidema (2009), Smith (2009), and Dediu (2009) show that the behavior of samplers that learn from multiple teachers do not converge to the prior (however, see Burkett and Griffiths (2010) for a treatment of iterated learning from multiple teachers that does yield convergence to the prior).

In addition to the computational modeling approaches described above, there have been empirical investigations into the problem of linkage (Kalish et al., 2007; Griffiths et al., 2008; Culbertson et al., 2012; Reali and Griffiths, 2009). All of these studies argue that the outcome of iterated learning reflects cognitive biases and they do so from three different approaches. The first approach, taken by Kalish et al. (2007) and Griffiths et al. (2008) shows that iterated learning "reveals inductive biases" by converging on a distribution over behaviors that mirrors known inductive biases. Previous psychological research has shown that people have an inductive bias favoring positive linear functions, because people tend to guess this relationship when exposed to little data and also learn this function faster than others (Busemeyer et al., 1997). Kalish et al. (2007) investigated this particular bias in iterated function learning experiment, in which participants were trained on two stimuli with lengths that were related to one another by a particular function: positive linear, negative linear, parabolic, or random. Participants were tested on the functional relationship they had learned and these testing responses were used as the training set for the next generation of learners. 32 iterated learning chains were initialized on one of these four functions (8 chains per function) and the data were passed down for 9 generations of learners. Behavior converged to a positive linear function in 28 out of the 32 chains, showing a strong bias for this function. Three chains (two in the negative linear condition and one in the parabolic condition) converged on a negative linear function, showing a weak bias for this function. The final chain did not converge. This finding of high rates of convergence upon the function that learners possess a bias for demonstrates the use of iterated learning as a tool for revealing such biases. Likewise, Griffiths et al. (2008) conducted an iterated category learning experiment on six types of category structures with known biases. Previous work by Shepard et al. (1961), corroborated by Nosofsky et al. (1994) and Feldman (2000), showed that Type 1 categories, where membership is defined along one dimension, is easiest for people to learn. Type 2 categories, where membership

is defined along two dimensions, is next easiest to learn. Type 3, 4, and 5, which have a single rule plus one exception proved harder, and equally difficult to learn. And Type 6, where no more than two members can share any given dimension is the most difficult to learn. Across 12 iterated learning chains, behavior converged to Type 1 and Type 2 categories, with Type 1 being more prevalent. These two experiments showed that the prevalence of behaviors that result from iterated learning conform with known inductive biases. This first approach, therefore, takes for granted that iterated learning is necessary to link biases to typological distribution, and treats convergence as an open empirical question: cultural transmission changes behavior over time, but do these changes converge upon cognitive biases or not?

A second approach treats the problem of linkage itself as an open empirical question and tests whether the biases evident in a *single* generation of learners are enough to mirror the known typological patterns in language. Culbertson et al. (2012) addressed this question by investigating four different word ordering patterns in nominals. Pattern 1 places adjectives and numerals before nouns, Pattern 2 places them after nouns, Pattern 3 is noun-adjective, numeral-noun and Pattern 4 is adjective-noun, noun-numeral. These four typological patterns are described in Greenberg (1963) as Universal 18 and have different prevalence in the world's languages. Pattern 2 is the most common (at 52%) , followed by 1 (at 27%), 3 (at 17%), and then 4 (at 4%). Culbertson et al. (2012) trained participants on an artificial language consisting of a mixture of these four possible orderings in four conditions plus one control condition. In the control condition, participants were trained on each ordering (adj-noun, num-noun, noun-adj, noun-num) in equal frequencies (25%, 25%, 25%, 25%), In condition 1, the two orderings consistent with Pattern 1 occurred 70% each, and the other two occurred 30% each (i.e. adj-noun, num-noun, noun-adj, noun-num: 70%, 70%, 30%, 30%, respectively). These same relative frequencies were used in the four conditions, but where condition 2 was compatible with Pattern 2, condition 3 compatible with Pattern 3, and condition 4 compatible with Pattern 4. In a testing phase, participants produced these orderings and the relative frequencies of their observed and produced orderings were compared. Participants did not reproduce the frequencies perfectly, showing evidence of a regularization bias (a bias toward over-producing one of the patterns, usually the most frequent one) and a substantive bias (a bias toward producing orderings that followed a particular pattern. The main result was that participants' substantive bias was ranked in rough conformance with the typological pattern: productions were biased most strongly toward Pattern 1 and Pattern 2, less strongly toward Pattern 3, and there was no evidence for a

bias toward Pattern 4. The authors state that the Pattern 1 bias may be over-represented in their data set because all of their participants spoke a Pattern 1 language, but overall they conclude that the biases of a single generation of learners mirror that of the typological pattern and thus, are a likely cause of this aspect of typology in the world's languages.

The third approach rests on the observation that Bayesian samplers' biases mirror typological distribution, whereas MAP learners do not, and address an open empirical question as to whether human learners are better approximated by samplers or MAP learners. As described in the beginning of Chapter 3, Reali and Griffiths (2009) conducted an iterated artificial language learning experiment where participants learned two names for six objects with different relative frequencies. After a few generations, learners regularized the language by eliminating one of the synonyms to create deterministic mappings between each object and its name. The intergenerational transitions among the relative frequencies of these words served as the data of interest and sampler and MAP models were fit to each data set. The sampler model provided a better fit than the MAP learner and this provides some support for claims that the mapping between cognitive biases and typological distribution is one-to-one.

In this chapter, I take a new approach to the problem of linkage, by focusing on the mapping between the biased behavior of a single generation of learners and the analytically-determined typological distribution of single generation data. Previous experiments have used experimental iterated learning chains to estimate the typological distribution, whereas I have used single-generation psychology experiments to estimate the Markov process that ultimately specifies the typological distribution. The following sections will provide an introduction to iterated learning as a Markov process and explain how the stationary distribution of this process relates to typological distribution. Then, I walk through the Markov process estimation procedure for the frequency learning data in Experiments 1 and 2 and show how these data can be represented as an empirical transition matrix. From each estimated transition matrix, both single-generation behavior (from any combination of observation ratios) and stationary behavior can be analytically determined and compared, without intervening sampling error obscuring their relationship. Finally, I will detail the nature of the mappings and discuss their implications for the predictability of culturally transmitted behavior on the basis of cognitive biases.

6.1 Extrapolation of data forward through cultural evolutionary time

6.1.1 Markov processes

In Chapter 1, cultural transmission was described in terms of iterated learning: a process in which the behavior produced by some individuals in the population serves as the input for other individuals. As individuals learn from one another, the behavioral repertoire of the population may change in response to the constraints that each learner imposes on their own productions. Simulations and experiments in iterated learning are often conducted on population structures in which each generation consists of one learner, and learners only receive input from the learner in the previous generation (Kirby, 2001; Brighton, 2002; Smith et al., 2003; Kirby et al., 2008; Smith and Wonnacott, 2010; Reali and Griffiths, 2009). This form of iterated learning is equivalent to a Markov process, and this equivalence provides a useful formal framework for the analysis of the dynamics of iterated learning (Griffiths and Kalish, 2007).

A Markov process is a discrete-time random process over a sequence of values of a random variable, $v_{t=1}, v_{t=2}, ..., v_{t=n}$, such that the random variable is determined only by its most recent value (Papoulis, 1984, p.535):

$$P(v_t|v_{t=1}, v_{t=2}, ..., v_{t-1}) = P(v_t|v_{t-1}) \tag{6.1}$$

This describes a memoryless, time-invariant, process in which the past values $(v_{t-2}, v_{t-3}, ..., v_{t=1})$ of the variable have no direct influence on the current value. This is the case for iterated learning chains when learners only observe the behaviors of their teacher, but not the behavior of their teacher's teacher, and so on. All of the possible values of the random variable constitute the state space of this system. A Markov process is fully specified by the probabilities with which each state will lead to every other state in the system and these probabilities between states can be represented as a transition matrix, $\mathbf{Q}$ (Norris, 2008, p.3).

Figure 6.1 shows an example transition matrix for a system with 11 states, $s_0, s_1, ..., s_{11}$, corresponding to different values of the random variable $v_1, v_2, ..., v_{11}$. Each cell in the matrix, $\mathbf{Q}_{ij}$, gives the transition probability from state $s_{i=t-1}$ to state $s_{j=t}$. The shading of the cells denote the transition probabilities between states. In an iterated learning process, a learner receives input that corresponds to a particular state (at time $t-1$), produces an output that corresponds to a particular state (at time t), and then that output state serves as the input state

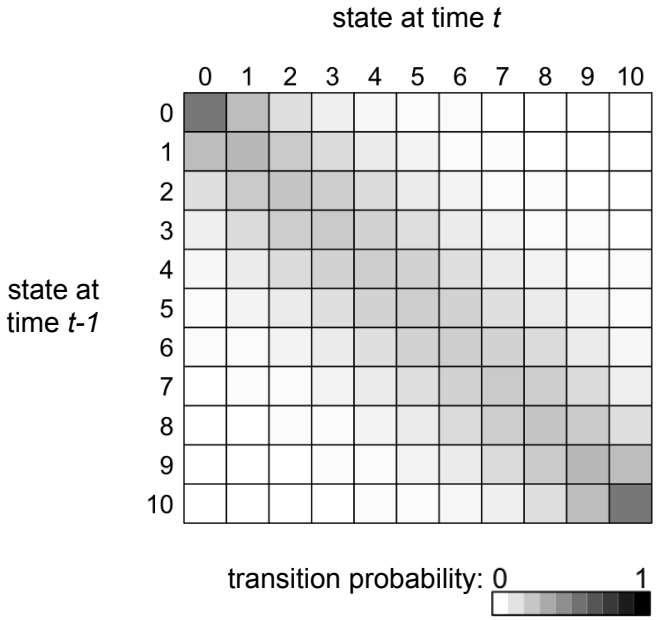


Figure 6.1: An example transition matrix for a system that has 11 states. Each state at time t-1 leads to each state at time t with a particular transition probability, coded in greyscale. White indicates a transition probability of zero and black indicates a transition probability of 1. For example, state zero (s_0) leads to s_0 52% of the time, s_1 26% of the time, s_2 12% of the time, and so on. Each row sums to one.

for the next learner. So not only does the transition matrix describe the behavior of individuals given every possible input, it also describes the behavior of an iterated learning chain across many generations of learners.

A transition matrix is the fingerprint of a system's dynamics and summarizes, in a sense, all of the forces that direct the evolution of the system. The particular transition matrix shown in Figure 6.1 is that of a Bayesian sampler with an unbiased prior (refer back to Chapter 5). Each component of the Bayesian model (the prior, the likelihood function, the hypothesis selection strategy, and the data production function) all contribute to the specific pattern of transition probabilities in the matrix. Likewise, if the learners here were human, the transition probabilities describing their behavior would be a summary of all of the cognitive biases at play in the inductive loop that links observations to their productions.

A transition matrix can also be used to address the problem of linkage for a particular system by determining the asymptotic behavior that the particular pattern of transitions give rise to. The asymptotic behavior is given by the stationary distribution over states and defines the amount of time that the system will spend in every state. This is the analogy to typological distribution that the

Bayesian iterated learning literature makes: typology is the distribution over cultural variants (i.e. states) after many generations of learners. The stationary distribution can be obtained from the transition matrix by different analytical and numerical methods (Stewart, 1994). One analytical solution is obtained by performing an eigen decomposition on the transition matrix. Here, the stationary distribution is proportional to the first eigenvector. A transition matrix will have one stationary distribution if it is *ergodic*, meaning that it is both irreducible (every state can be reached by every other state) and aperiodic (Griffiths and Kalish, 2007). Ergodicity is most often violated by the existence of sinks. For example, drift without mutation defines a transition matrix with sinks: once a variant is lost, it can never re-enter the population. In a population with only two variants, x and y, the system will either converge to a population state composed entirely of *variant x* or entirely of *variant y*.

One objective of cultural evolution experiments are to probe the veridical transition matrix in attempt to estimate the stationary distribution. If an iterated learning chain is initialized in an arbitrary state, and run for a long enough time, the frequency with which it passes through each state will converge toward the stationary distribution. The proportion of time the chain spends in each state will provide an estimate of the stationary distribution. Additionally, the transition matrix itself can be estimated by traditional experimental paradigms in which the different input states serve as different training conditions, and the distribution of participant responses provides an estimate of the corresponding row in the transition matrix. When the state space of the system is enumerable and small, estimating the transition matrix in this way may be a more efficient use of participants, but when the state space it large, iterated learning may be better. However, the relative efficacy of each of these estimation techniques is an open empirical question awaiting a solution.

6.1.2 Empirical transition matrices

Experiments 1 and 2 in this thesis were designed so that the data collected would constitute an estimate of the veridical transition matrix. In this section, I walk through the transition matrix representation of the data from these experiments.

Figure 6.2 provides the raw data from Experiment 1 in transition matrix format. Recall that in this experiment, participants observed blue and orange marbles being drawn from a bag in a particular ratio. Because participants observed 10 draws, there are 11 possible ratios of blue to orange marbles: 0:10, 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1, 10:0. These ratios constitute the 11

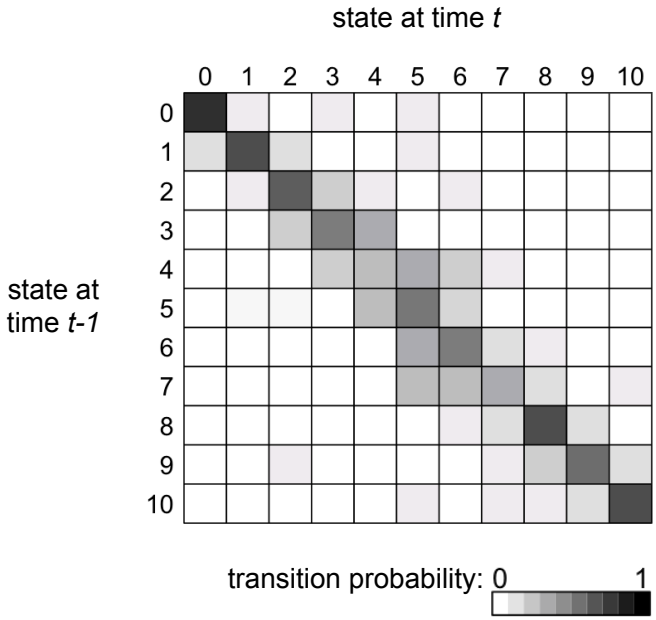


Figure 6.2: The data from Experiment 1, represented as an empirical transition matrix. Each state is defined by the number of blue marbles in the observation phase (at time $t-1$) or the production phase (at time t). Each state at time $t-1$ leads to each state at time t with a particular transition probability, coded in greyscale. Each row corresponds to one observation ratio. The transition probability equals the percentage of participants that responded with a particular production count of the blue marble (s_t), given a particular observation count (s_{t-1}). White indicates a transition probability of zero and black indicates a transition probability of 1. For example, state one (s_1) leads to s_0 13% of the time, s_1 69% of the time, s_2 13% of the time, and so on. This means that participants who observed 1 blue marble and 9 orange marbles produced no blue marbles 13% of the time, one blue marble 69% of the time, and 2 blue marbles 13% of the time. Each row sums to one.

possible states that this system can be in and are represented in Figure 6.2 by the count of the blue marble draws. States at time $t-1$ give the blue marble count in a participant's observations, and states at time t give the blue marble count in a participant's productions. Each row gives the distribution of participant responses per observation ratio condition. The rows labelled 0 to 4 and 6 to 10 each contain data from 16 participants, and the row labelled 5 contains 32 participants. The transition probabilities are taken directly from the proportion of participants in each condition that responded with each blue marble count.

Assuming that these raw data are a good estimate of the transition matrix, we can work out the distribution of responses that would be obtained from this experiment given any combination of training ratios. This is because the transition matrix carries complete information about all possible behavior. If all

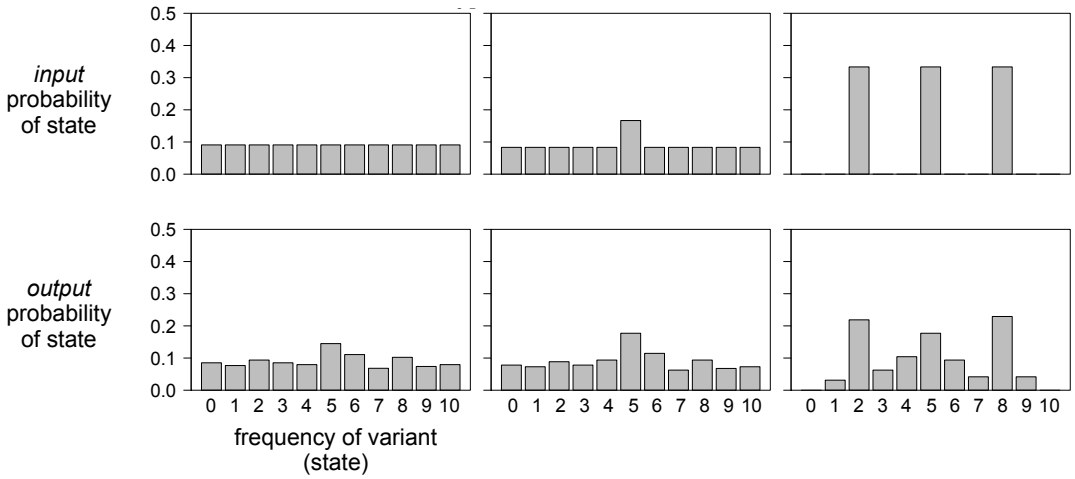


Figure 6.3: Top row) three hypothetical distributions over observation ratios. Each observation ratio is shown on the x-axis in terms of the count of the blue marble variant. Bottom row) the distribution over production ratios for Experiment 1, that would be obtained from each input distributions above. The top right pane shows an input distribution of a typical psychology experiment. Here, only three input states are tested with, for example, 32 participants trained on a 2:8 ratio, 32 participants trained on a 5:5 ratio, and 32 participants trained on an 8:2 ratio. The bottom left pane shows what the results of that experiment would look like. This distribution provides the worst estimate of the stationary distribution for Experiment 1.

participants were given a uniform distribution of observation ratios, then the distribution of over all participant responses are given by the marginal distribution (summing over columns). But if a non-uniform set of input states are used, then the marginal distribution weighted by this input distribution yields the distribution of participant responses. Figure 6.3 shows some example input distributions (top) and their resulting output distributions (bottom). The two middle panels show the actual input distribution used for Experiment 1 and the resulting distribution of participant responses.

This empirical transition matrix provides all of the information needed to predict the cultural evolution of this system. The idea of culturally transmitting marbles in bags seems odd if taken literally, but it provides a good example of how probability matching behavior leads data sets to evolve over time. If Experiment 1 were an iterated learning experiment, the first participant (g_0) in the chain would be shown a particular observation ratio, say 5:5 (corresponding to s_5). If g_0 produced a 6:4 ratio (s_6), then the next participant (g_1) would be shown a 6:4 ratio, and so on. Figure 6.4a shows a simulated run of an iterated learning chain

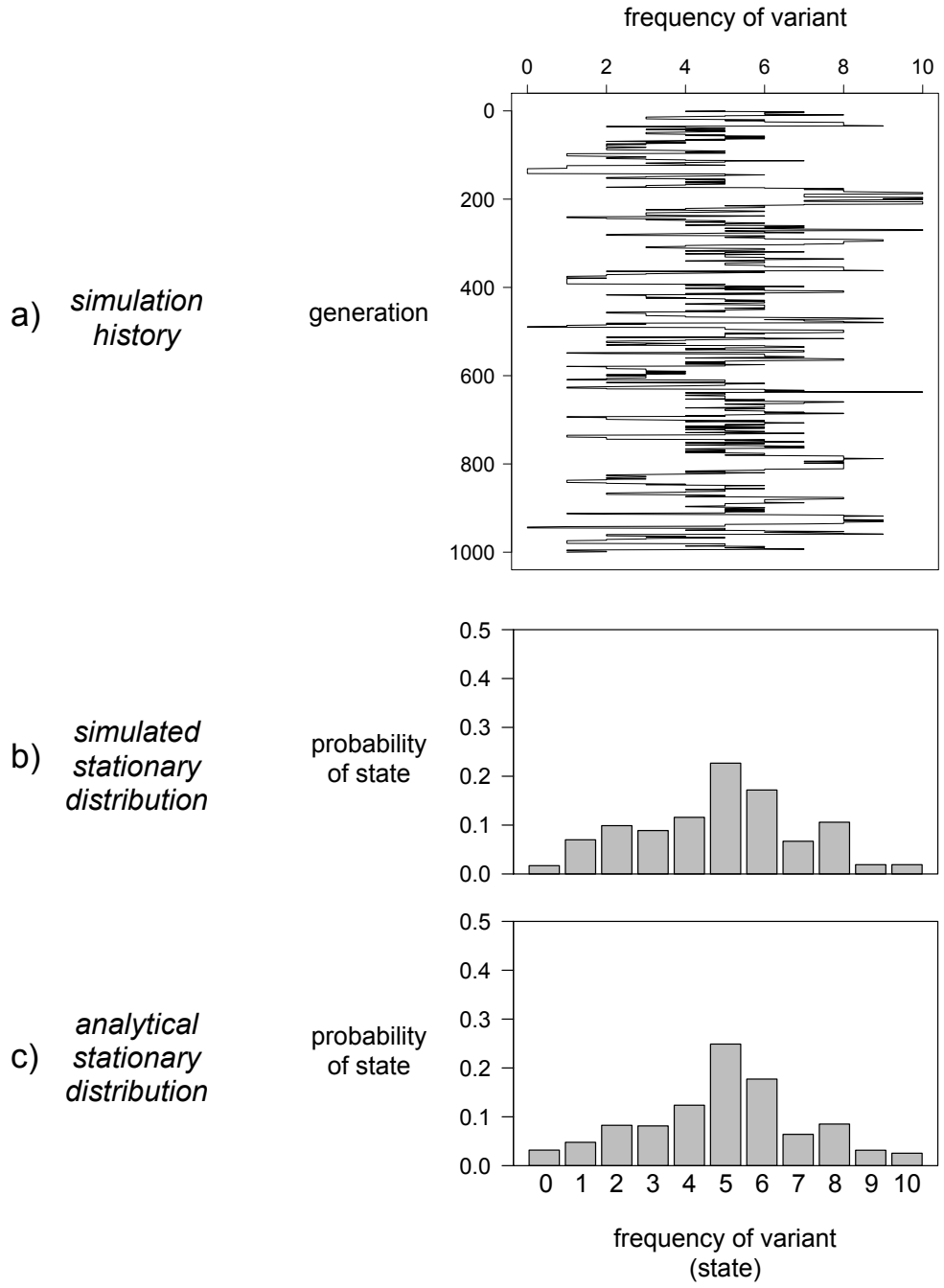


Figure 6.4: a) A simulated iterated learning chain using the transition probabilities between the raw data in Experiment 1 (i.e. the raw empirical transition matrix in Figure 6.2). b) Probability distribution over the amount of time (in generations) that the simulated chain spent in each state. This provides an estimate of the stationary distribution. c) The analytical solution to the stationary distribution, calculated by eigen decomposition from the raw transition probabilities in Experiment 1. The variant being tracked throughout is the blue marble.

on this raw empirical transition matrix, initialized at s_5, for 1000 generations. Figure 6.4b shows the distribution of states that this chain visited, which should approximate the stationary distribution of this system. Figure 6.4c shows the analytical solution of this system's stationary distribution, calculated via eigen decomposition of the raw empirical transition matrix. This 1000-generation chain provides a good approximation of this matrix's stationary distribution. What we learn from the stationary distribution, is that the system is biased toward 5:5 ratios, because it spends most of its time in this state. This corroborates the Bayesian model fitting results in Chapter 5 that returned a best-fit prior bias toward variability in Experiment 1. If the system were unbiased, and participants were probability matching in a stricter sense, the stationary distribution would be uniform. And if the system were regularizing, the stationary distribution would place most mass on the outer edges (ratios 0:10 and 10:0). The stationary distribution provides us with a different understanding of the data at hand. Although participants in Experiment 1 were probability matching (the mean of responses was not significantly different from each input frequency) the pathways formed by the transition probabilities funnel the system toward the 5:5 state. This represents a bias in participant behavior that is more subtle than the one understood by analyzing the distribution of responses per input ratio *in isolation from one another*, and exemplifies what is meant in the literature by "iterated learning reveals inductive biases" (e.g. Kalish et al., 2007; Griffiths et al., 2008).

The analyses presented above dealt with raw data, but there are a variety of ways that the estimate of the transition matrix can be improved. The first one involves smoothing the data. This adds a small, uniform level of noise to the matrix and renormalizes it. In the remaining analyses in this chapter, I will use a Grassberger prior in which a pseudo count of $\frac{1}{length(row)^2}$ is added to each cell in the matrix (Grassberger, 2003; Wolpert and DeDeo, 2013). Second, because regularization concerns biases that operate on relative frequencies, and the analyses have shown no significant effect of the specific stimuli on production ratios, we can use the data from all stimuli to estimate how participants produce the majority marble across the different observation ratios. This is exactly analogous to what would be done in a typical iterated learning experiment, where stimuli are randomized between generations and only the evolving property of interest (i.e. the relative frequency of the variants) is preserved and passed between participants (e.g. Kirby et al., 2008; Reali and Griffiths, 2009; Smith and Wonnacott, 2010). Therefore, I will collapse the data over marble colors and word types as done in the previous chapters. This yields a 6×11 matrix covering states s_6 through s_{10} only. Because transition matrices need to be square for most matrix

operations, including eigen decomposition, these data will be copied into states s_0 through s_4 to yield an 11×11 matrix. The states will now be defined by the count of an arbitrary marble or word, where s_0 through s_4 shows what happens to minority variants and s_6 through s_{10} shows what happens to majority variants. Figure 6.5 and 6.6 shows the empirical transition matrices, estimated in this way, for the four conditions in Experiment 2. Each matrix's marginal and stationary distributions are also displayed and will be discussed in the next section.

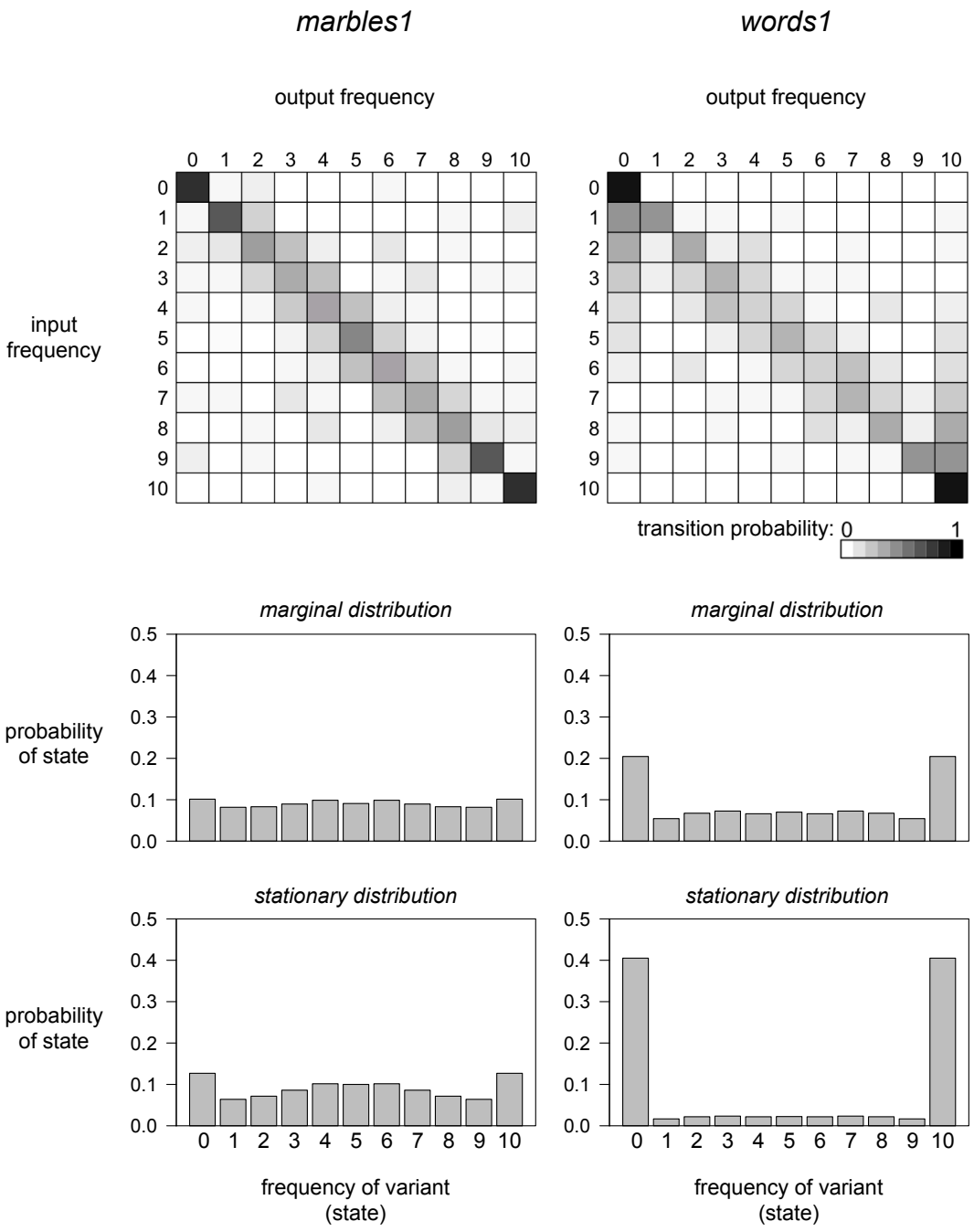


Figure 6.5: Empirical transition matrices (top) for the one-item conditions in Experiment 2 (*marbles1* and *words1*) and the marginal (middle) and stationary (bottom) distributions calculated from each matrix.

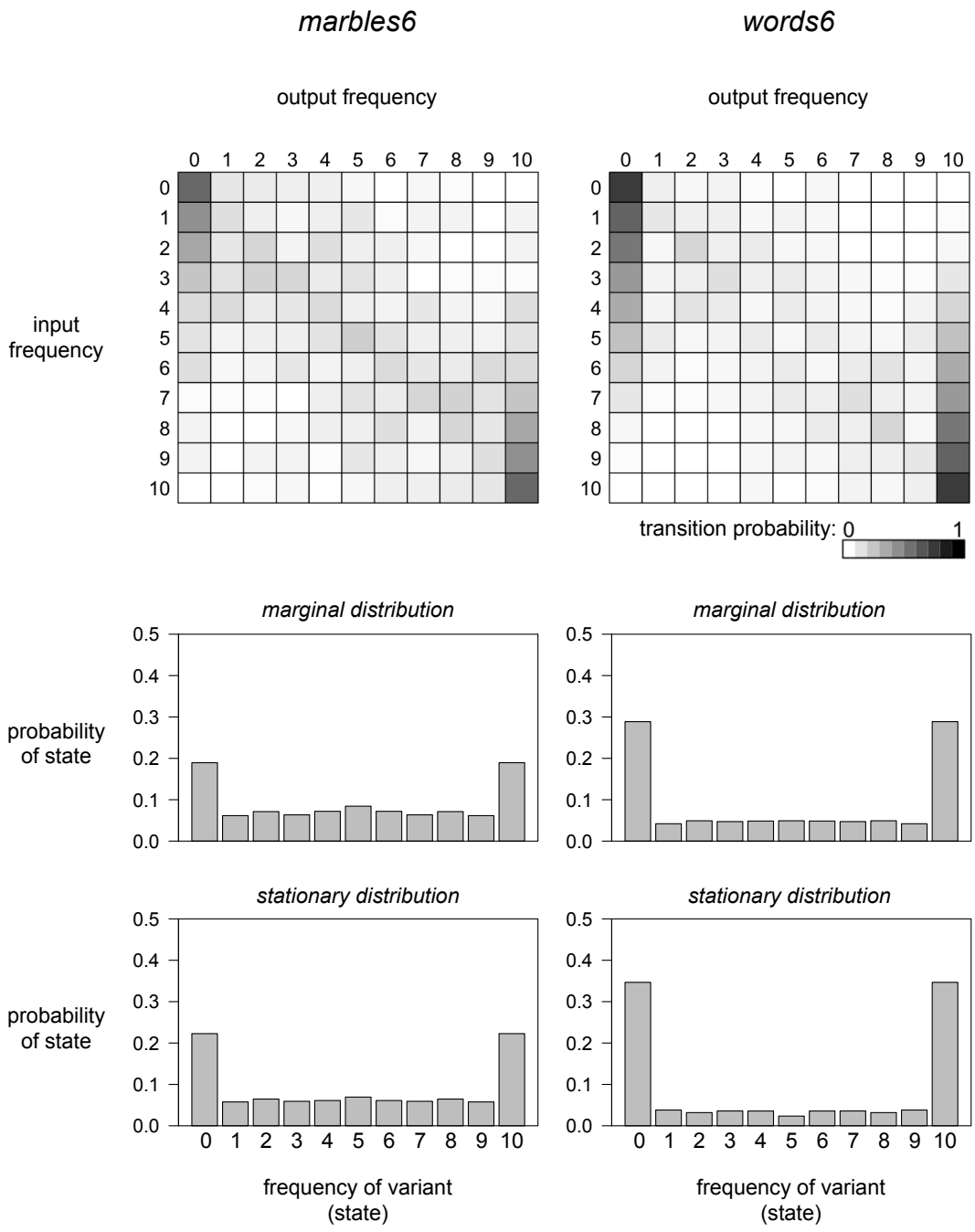


Figure 6.6: Empirical transition matrices (top) for the six-item conditions in Experiment 2 (*marbles6* and *words6*) and the marginal (middle) and stationary (bottom) distributions calculated from each matrix.

6.1.3 Eigen decomposition of participant behavior

In this section, I investigate how the distribution of participant behavior will evolve over several generations of learners. More precisely, what are the paths of change that connect any given input distribution to the stationary distribution?

Let us begin by focusing on the *marbles1* and *words1* data sets. In both of these conditions, the population of participants were trained on a uniform distribution over input ratios {5:5, 6:4, 7:3, 8:2, 9:1, 10:0} and the resulting distribution differed markedly as a result of participants' biases (refer back to Figure 3.12). The marginal distributions in Figure 6.5 tell us what participant behavior will look like after one generation of learners. But how will this distribution develop after a second generation of learners? This question could be answered with an experimental iterated learning approach where we train another population of 192 participants on the distribution produced by generation 1. However, we already have a good estimate on how participants will respond to each observation ratio in the state space, so there is no need to run this experiment again.

One way to get an idea of what this second generation distribution will look like is by sampling it from the new input distribution, with the probabilities given in the transition matrix. To improve the picture, these resamplings can be averaged. However, the veridical average of these resamplings can be obtained analytically via some basic matrix algebra, by multiplying g_1's output distribution by the transition matrix to yield g_2's output distribution. This process can be repeated to obtain g_3's distribution on the basis of g_2's output distribution, and so on. It can also be solved for an arbitrary pair of generations. Let the output distribution of g_n be $g_{n+1} = \vec{v}_{n+1}$. Then, $\vec{v}_{n+1} = \vec{v}_0\mathbf{Q}^n$, where $\vec{v}_0$ is the initial distribution over observation ratios. It is also worth mentioning here that the defining feature of the stationary distribution, which is the vector that yields itself when multiplied by the transition matrix: $\vec{s}\mathbf{Q}= \vec{s}$, where $\vec{s}$ is the stationary distribution.

Figure 6.7 shows how a uniform initial distribution will develop over several generations of learners in the linguistic and non-linguistic domains during single frequency learning. In generation 1, we see the uniform distribution characteristic of unbiased probability matching behavior in the non-linguistic domain, and a clear bias for fully-regular mappings in the linguistic domain. Over time, the linguistic regularization bias leads to more and more regularity in the system, leveling off at 40% 0:10 ratios and 40% 10:0 ratios. Strikingly, the *marbles1* data gradually develops some regularity over time. This stationary distribution is quite different from that obtained in Experiment 1 (Figure 6.4c) for a nearly identical

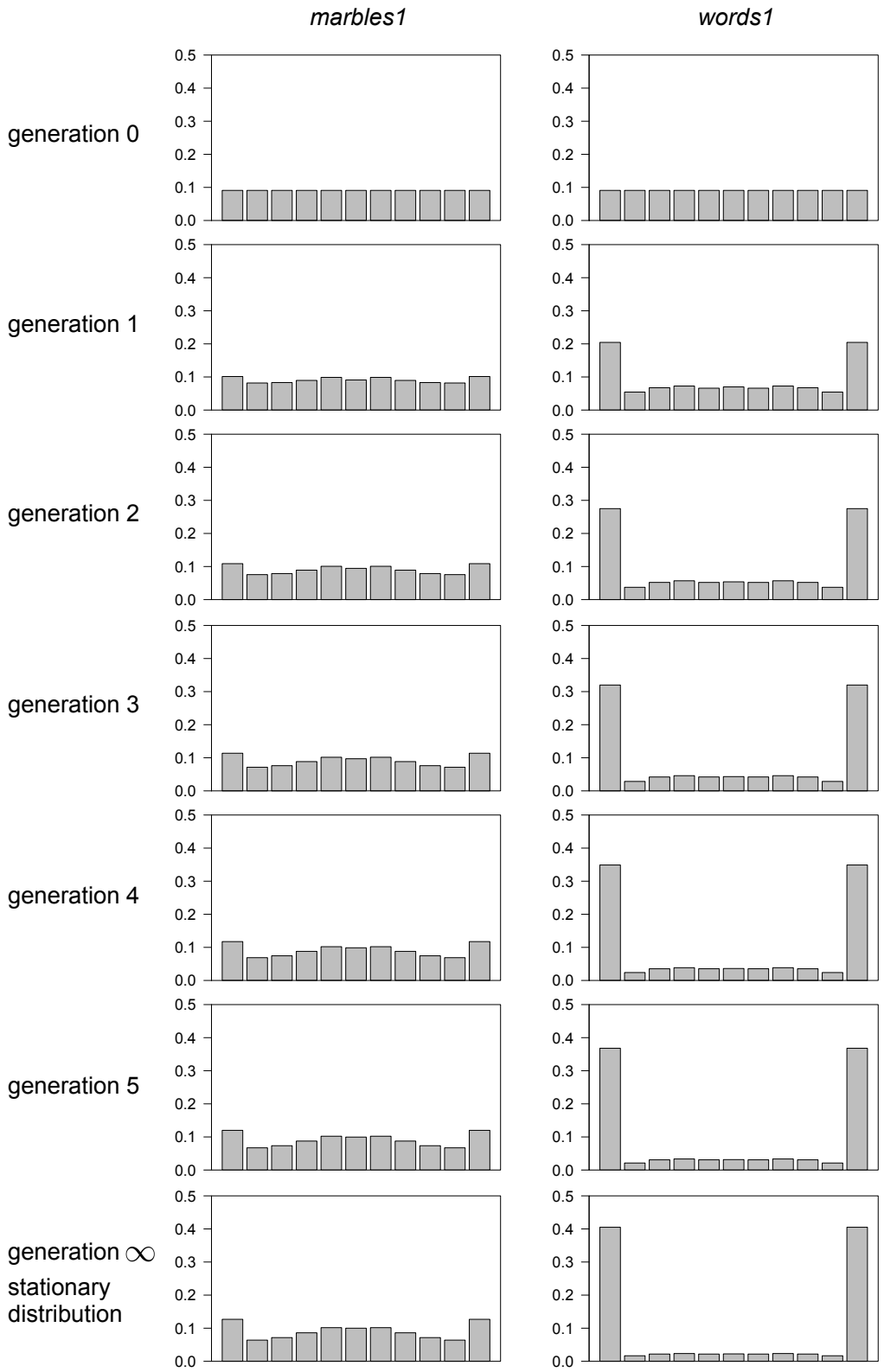


Figure 6.7: The evolution of a population of participants' behavior during non-linguistic (*marbles1*) and linguistic (*words1*) frequency learning, as calculated from empirical transition matrices.

experimental design.[2] Although both of these experiments elicited probability matching behavior, the specific pattern of transition probabilities funnels behavior in Experiment 1 toward 5:5 ratios and funnels behavior in *marbles1* toward 0:10 and 10:0 ratios. This exemplifies the fact that apparently similar behaviors can have very different evolutionary consequences.

Figure 6.6 shows the marginal and stationary distributions for *marbles6* and *words6*. It is important to note here that in order to describe the six-item data with an 11-state Markov process, the transition probabilities associated with each input state must be independent of the other input states. However, this was not the case for the two six-item conditions because each ratio was learned in the context of five other ratios. In an iterated version of the six-item conditions, the set of six observation ratios would change each generation. The data collected in *marbles6* and *words6* only tested 3 of the $6^6 = 46656$ possible observation ratio sets (refer back to Experiment 3). However, on the basis of the data from these three observation ratio sets, I have reason to believe that the these estimates are adequate, because the distributions of the *all 5:5* observation set was not significantly different from that of the 5:5 ratio when learned in the context of other ratios. However, the distributions of the *all 10:0* observation set was significantly different than the 10:0 ratio when learned in the context of other ratios. So perhaps the transition probability estimates for the {5:5, 6:4, 7:3, 8:2, 9:1} ratios are fair, but the 10:0 ratio underestimates the extent to which this area of the transition matrix forms a sink. One way to validate these stationary distribution estimates would be to re-run the experiment with a distribution of observation ratios that equals the estimated stationary distribution. If the resulting distribution is the same, then this would be evidence that this is the stationary distribution (based on the definition of the stationary distribution as the vector that yields itself when multiplied by the transition matrix).

Compared to the one-item tasks (Figure 6.5), the marginal distributions of the six-item tasks are much more similar to the corresponding stationary distributions (Figure 6.6). This means that behavior due to concurrent frequency learning biases converge to the stationary distribution in fewer generations than the domain-specific biases alone. This convergence rate can also be obtained from the eigen decomposition of the transition matrix. When the second eigenvalue, λ_2, is closer to 1, the influence of the start state remains longer, meaning that

[2]The only differences were: 1) Experiment 1 only used orange and blue marbles, whereas *marbles1* tested one of the 6 color pairs (used in *marbles6*) at random. 2) Experiment 1 counterbalanced the test side location of the marble colors between participants, whereas *marbles1* randomized the test side location of the marble colors per test trial.

it will take more generations for an iterated learning chain to approximate the stationary distribution (Rosenthal, 1995). Of the four experimental conditions, *marbles6* will converge to the stationary distribution fastest, with $\lambda_2 = 0.7$, then *marbles1* with $\lambda_2 = 0.81$, then *words6* with $\lambda_2 = 0.84$, and finally *words1* with $\lambda_2 = 0.92$.

This means that the behavior of a single generation in the non-linguistic conditions carries more information about the stationary distribution than those of the linguistic conditions. Another way of making this point is by looking directly at the amount of information each marginal distribution carries about the stationary distribution. Kullback-Liebler (KL) divergence is one such information-theoretic distance measure and can be used to quantify the additional amount of information needed to specify the marginal distribution, above and beyond the information provided by the stationary distribution. If zero additional bits are needed, then this means that the marginal distribution carries full information about the stationary distribution (i.e. the typological distribution of the system can be perfectly predicted from the behavior of one generation of learners). So, the lower the KL divergence, the more information the marginal distribution carries about the stationary distribution. KL divergence scores mirror the convergence rates for the four conditions, with the marginal distribution of *marbles6* carrying the most information about its stationary distribution at 0.014 bits, then *marbles1* at 0.018 bits, *words6* at 0.053 bits, and *words1* at 0.566 bits.

6.2 The obscure mapping of biases to behavior

When we ask how cognitive biases map onto the typological distribution of languages in the world, we are asking about two types of mappings: 1) how do cognitive biases map on to individual behavior and 2) how do individual behaviors map onto the structure of languages. In the previous chapters, I have addressed the first question and shown that three different kinds of biases regarding linguistic stimuli, concurrency, and coordination, all lead to similar amounts of regularization in a basic frequency learning task. This demonstrates that the behavioral profiles of individuals can underspecify the particular cognitive biases at play: regularization behavior is multiply realizable. This obscures the mapping of cognitive biases to single generation behavior.

This chapter has focused on the second question. First we have seen that the biases of one generation of learners does not yield a distribution over behavior that mirrors the stationary distribution. This proves that the cultural transmission process is an integral part of the problem of linkage: it is the mechanism that links individual behavior to typological patterns. The functionalist interpretation exemplified by Culbertson et al. (2012) may hold for certain combinations of biases and typological patterns. However, it should not be taken as supporting evidence that biases will necessarily mirror typological patterns.

In addition to not finding a one-to-one correspondence between distributions, the coarser-grained characterization of bias to typological distribution mirroring in terms of *ranking* preservation does not seem to hold either. For each condition in Experiment 2, I calculated the regularity of participant behavior (on the basis of a uniform input distribution to the estimated transition matrix) for one generation of learners and compared this to the regularity level achieved in the stationary distribution of that matrix. Figure 6.8a shows these regularity levels, given on the y-axis in terms of entropy, according to the measure of regularization described in Section 3.1.[3]. The regularity of participant productions are, in order of increasing regularity: 0.64 bits for *marbles1*, 0.50 bits for *marbles6*, 0.47 bits for *words1*, and 0.34 bits for *words6*. The regularity levels of the stationary behavior were also calculated by the same measure and are, in order of increasing regularity: 0.61 bits for *marbles1*, 0.44 bits for *marbles6*, 0.24 bits for *words6*, and 0.15 bits for *words1*. Here we see that the ranking is not preserved. The cognitive biases associated with particular levels of regularity in a single generation of learners do not map on to the levels of regularity after cultural evolution takes place. Therefore, the

[3]Recall that this calculation of entropy is identical to both the average entropy of each synonyms set, and the conditional entropy of the entropy of the entire set of production ratios.

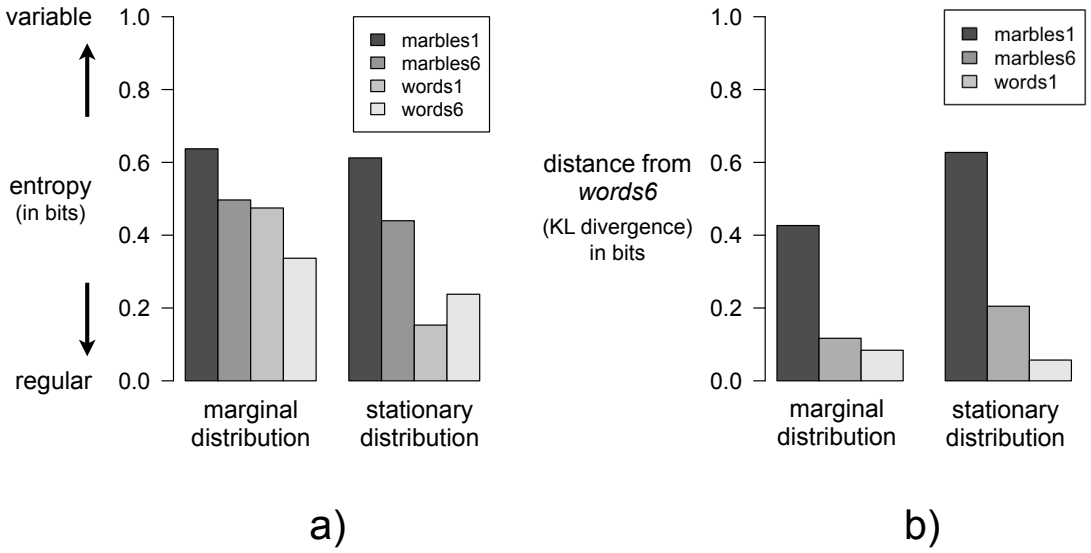


Figure 6.8: a) Regularity of single-generation and stationary behavior in the four conditions of Experiment 2. b) KL divergence of participant behavior in *marbles1*, *marbles6*, and *words1* from participant behavior in *marbles6*.

mapping between single generation and stationary behavior is obscure as well.

Finally, to return to one of the main questions explored in Experiment 2: what are the relative contributions of domain-specific and domain-general biases to the linguistic regularization bias? In Chapter 3, this question was addressed by taking a detailed look at the predictors of regularization behavior within a single generation of learners. Now that we have access to the stationary distribution, we can ask how much information participant behavior in each condition carries about participant behavior in the full artificial language learning task *words6*, and whether one of the factors becomes more predictive over several generations of learners. Again, information distance is calculated using Kullbach-Liebler divergence. Figure 6.8b shows the distance of each condition (*marbles1*, *marbles6*, and *words1*) from *words6* among the marginal distributions, and among the stationary distributions. As expected, *marbles1* behavior carries the least amount of information about *words6* behavior. What is of interest here is whether multiple-frequency learning or linguistic domain carries more information about the regularity of language. This proves to be the linguistic domain, both when comparing single-generation behavior, and even more so when comparing stationary behavior. On the basis of these estimated transition matrices, multiple-frequency learning and linguistic domain carry information about linguistic regularity, but domain-specific biases seem to be more predictive of the level or regularity in

language.[4]

This kind of directional change over time was also exemplified by the striking difference between stationary distributions in the two probability matching profiles of Experiment 1 and *marbles1*. Neither of these experiments elicited a change in entropy that was significantly different from zero, with marginal distribution entropy of 0.68 bits for Experiment 1 and 0.64 bits for *marbles1* (input entropy was 0.64 bits in both cases). However, their stationary levels or regularity differ markedly: 0.83 bits for Experiment 1, representing a strong increase in variability over time, and 0.61 bits for *marbles1*, representing a slight increase in regularity over time.

In summary, cultural transmission is an integral part of the problem of linkage. In none of these experiments did the distribution of single-generation behavior mirror that of the stationary behavior. In all cases, the distribution of behavior changed over generations, becoming more regular over time. Furthermore, the relative gains in regularity differed between conditions and lead to different rankings of regularity within the marginal and stationary distributions. Differences in the pattern of transition probabilities lead to different stationary levels of regularity and nuanced biases not readily seen in single generation behavior. Therefore, these stationary levels of regularity would be difficult to predict on the basis of single-generation behavior that only experimentally tests a small subset of the input states. Such results would necessarily be over-fit to participants' behavior given those particular inputs and contain very little information about the likelihood that those behaviors would lead to other behaviors for any scale of time greater than one generation. This is because intermediate, untested, input states may provide unexpectedly high-probability or unexpectedly low-probability transitions to other states that were tested. Information about these probabilities simply can not be ascertained from investigating a sparse sample of all possible input states. One piece of advice to experimentalists who are attempting to obtain the stationary distribution from a single generation of learners would be, once you think you have found the stationary distribution, train another population of participants on that distribution and see if the population returns the same distribution. If so, then this is probably the stationary distribution. How-

[4]From personal communication with Tom Griffiths, there do not seem to be any existing methods for obtaining confidence intervals on stationary distributions derived from estimated transition matrices. It is unknown how the error associated with an estimated transition matrix is related to the error on the stationary distribution when calculated from that matrix. This same problem also holds for iterated learning experiments which estimate the stationary distribution on the basis of several trajectories through the veridical transition matrix. This is an important area for future methodological development, but for now I will have to discuss these results without confidence intervals.

ever, if the state space is large, this type of verification will be difficult and an iterated-learning based design that initializes chains on a broad sample of start states may be the next best method. In short, psychological experimentation is a powerful way to relate cognitive biases to behavioral biases, but requires special attention to experimental design when the object of inquiry is the structure of behaviors that emerge as the result of cultural evolution.

6.3 Cognitive biases as selection pressures

This thesis has explored how the human mind alters the distribution of variation in culturally transmitted data sets. The story line I have put forward has framed cognitive biases as analogical to selection pressures on culture as it evolves. In particular, I have explored a variety of basic biases in frequency learning and production and in this section I would like to clarify the relationship between these biases and certain selection functions in the genetic and cultural selection literature.

In Chapter 1, I reviewed two genetic models of selection and frequency-dependent selection and two cultural models of direct bias and frequency-dependent copying. Then, I presented several experiments that targeted cognitive biases during frequency learning and therefore, the cognitive basis of frequency-dependent copying. By design, these experiments utilized stimuli that were not likely to be selectively copied on the basis of their intrinsic properties (such as a color of the marbles stimuli or the phonetic or orthographic realization of the word stimuli) and the data analyses verified that these properties did not effect participants' regularization behavior. The only exception was in Experiment 5 on tacit coordination, where participants successfully coordinated on marble color, but only in the 5:5 condition where the marbles could not be distinguished in terms of relative frequency to one another. In the 7:3 condition, participants coordinated on the majority marble instead of a particular marble color. Overall, it appears that these weak but potential sources of direct bias are either washed out by frequency-based selection pressures, or completely ignored when frequency-based copying is possible. This suggests that frequency-based selection pressures may be more common in cultural evolution than genetic evolution, but this is only a tentative suggestion which implies a variety of interesting follow-up experiments. Additionally, the Bayesian models of frequency learning should constitute a form of frequency-dependent selection. Because there was no evidence for direct bias in the experimental data to which these models were fit, the set of models were

decidedly restricted to those which are incapable of producing direct bias: those with symmetrical priors (refer back to Section 5.1 and Figures 5.1 and 5.2 for the explanation of this). Therefore, in the remainder of this section, I will discuss the similarities and differences between 1) the selection functions defined by the human frequency learning data and the Bayesian models of frequency estimation and production and 2) the models of frequency-dependent selection and frequency-dependent copying described in Chapter 1.

Figure 6.9 plots some example behavior of the Bayesian models in the format of a selection function (refer back to the plots in Section 1.6). All of the functions defined by this model can be understood as forms of frequency-dependent selection, because they define values of θ' as a function of the frequency of θ.[5] Although, none of these Bayesian models map on directly to the frequency-dependent models of Felsenstein (Figure 1.9) and Boyd and Richerson (Figure 1.12) described in Section 1.6. In Boyd and Richerson's terms, the maximizer model can produce both conformity and anti-conformity copying, however the sampler and averager models can only produce anti-conformity copying. The top left pane of Figure 6.9 constitutes a conformity copying function for maximizer with a regular prior of $\frac{\alpha}{2} = 0.1$. Here we see that the more frequent a variant is, the more likely it is to be copied. This leads to the selective elimination of variation which would cause the entropy of agent productions to drop each generation (on average), and constitutes regularization. For all values of $\frac{\alpha}{2} < 1$, the maximizer model constitutes conformity copying. As expected, when the prior is unbiased ($\frac{\alpha}{2} = 1$) the maximizer model achieves drift and is identical to Boyd and Richerson's frequency-dependent copying model when $D = 0$. And for all priors that bias agents toward variability ($\frac{\alpha}{2} > 1$), the maximizer model defines anti-conformity copying. This leads to the selective maintenance of rare variants in the population and does not constitute a regularization process. The second row of Figure 6.9 shows some selection functions from the sampler and averager models. When $\frac{\alpha}{2} = 0$, these models define drift and are identical to Boyd and Richerson's frequency-dependent copying model when $D = 0$. For all other values of $\frac{\alpha}{2}$, these models define anti-conformity copying. One of the main mechanisms behind a class of selection known as *balancing selection* is frequency-dependent selection for rare variants (e.g. Clarke, 1979; Maynard Smith, 1989; Mokkonen et al., 2011). Balancing selection is any form of selection that maintains variation in a population and was originally conceived to account for balanced polymorphism in the population genetics literature (e.g. Dobzhansky, 1951; Wright, 1969)

[5] but see Reali and Griffiths (2010) for a proof that the averager model can be described as Wright-Fisher drift with mutation.

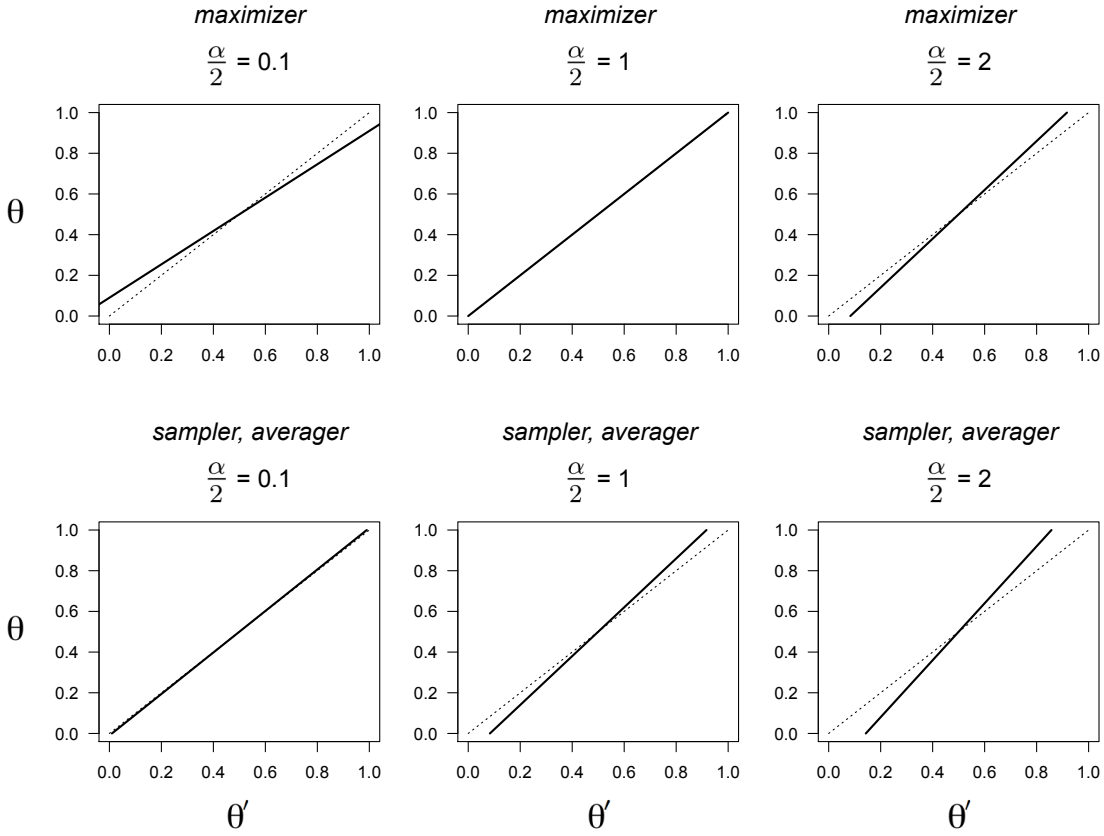


Figure 6.9: Some example selection functions as defined by the Bayesian models for different values of the prior parameter $\frac{\alpha}{2}$, where $N = 10$. This function maps θ (the input proportion of *variant* x at generation $t-1$) to θ' (the expected proportion of *variant* x at generation t). θ' is simply the mean of the model's behavior for a given input proportion (i.e. the mean of each distribution as shown in Figures 5.7, 5.8, and 5.9). The sampling error about the mean, which differs between the models, is not depicted in this plot. The sampler and averager models are identical in terms of the means they produce per $\frac{\alpha}{2}$, but they differ in terms of the sampling error about this mean. The maximizer model achieves frequency-dependent selection for the common variant when $\frac{\alpha}{2} < 1$, it is neutral and defines drift when $\frac{\alpha}{2} = 1$, and it achieves frequency-dependent selection for the rare variant when $\frac{\alpha}{2} > 1$. The sampler and averager models are only capable of frequency-dependent selection for the rare variant, except when $\frac{\alpha}{2} = 0$ and these models define drift.

In extreme cases, balancing selection achieves a uniform distribution over variants in a population, and this relates closely to the variability bias in these Bayesian models. As $\frac{\alpha}{2}$ approaches infinity, model behavior is increasingly biased toward a 50/50 production of the two variant types. In the Dirichlet-multinomial version of this beta-binomial Bayesian model, a uniform distribution over all variants is achieved when $\frac{\alpha}{2}$ goes to infinity. Balancing selection can not, however, stabilize variants at non-50/50 ratios in the way that probability matching behavior can: it only relates to the variability bias and not to any cognitive mechanisms that lead to probability matching and the high transmission fidelity of particular ratios of variants. As models of regularization, however, the sampler and averager are unsatisfactory (at best) and confusing (at worst). These models do possess a prior that can be biased toward regularity, but they are incapable of producing regularization via conformity copying behavior (refer back to p. 179 for a discussion of how the sampler achieves a limited amount regularization behavior by way of its broad sampling error). Although these models seem inadequate for describing human regularization behavior, they do constitute two types of the many theoretically possible models of inductive evolution.

Figure 6.10 plots the selection functions defined by the raw data in Experiment 2. Here, θ is the observed proportion of *variant x* and θ' is the mean proportion of *variant x* that participants produce. Compare these plots to the transition matrices in Figures 6.5 and 6.6: the selection function connects the means of each transition matrix row[6] and omits the sampling error. Strikingly, all of the plots in Figure 6.10 appear to define noisy renditions of drift[7], but we know from the detailed analysis of the complete distribution of responses, that participants are most certainly regularizing in *marbles6*, *words1*, and *words6*. This difference is due to the unique spread of error about the mean. Participants regularize by overproducing one of the two variants, yielding a bimodal distribution with a mean that tends to equal the observation proportion. This achieves regularization in a very different way than Felsenstein or Boyd and Richerson's models of frequency-dependent selection do: these models assume binomial error about the mean, but participant responses do not conform to this, and this leads to different evolutionary dynamics. If participants only regularized by overproducing the majority variant, then the resulting selection function would look much more like these basic models of frequency-dependent selection. However, some participants regularize by overproducing the minority variant, in such a way that

[6] please pardon the inverted y-axis...

[7] though *marbles1* and *marbles6* could be a frequency-dependent copying function that favors the minority variant.

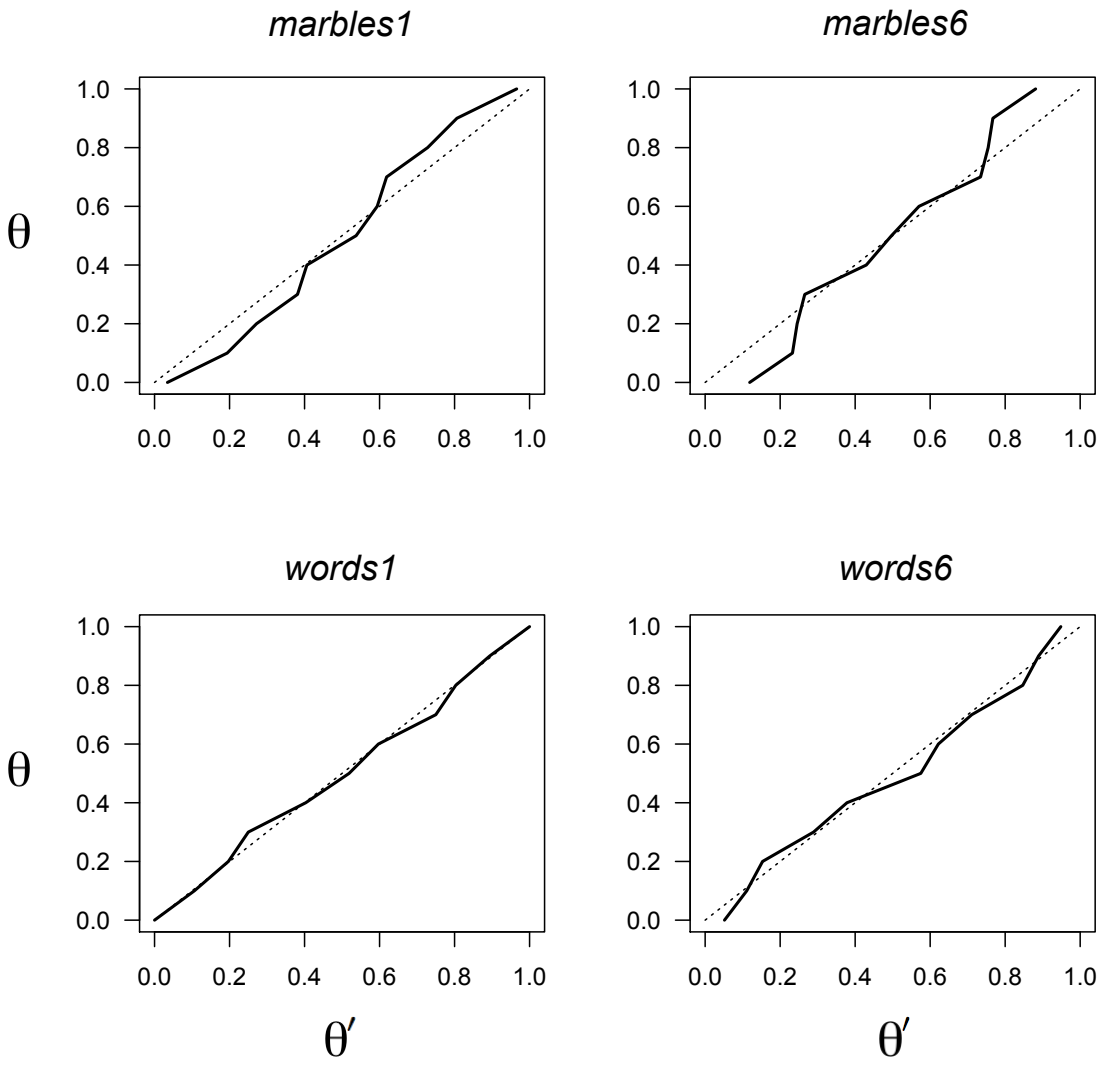


Figure 6.10: Participant data from Experiment 2, plotted as a selection function which maps θ (the observed proportion of *variant x*) to θ' (the mean production proportion of *variant x*. The sampling error about the mean is not depicted in this plot.

the population mean falls on input frequency, even though all participants were regularizing. Whereas the Felsenstein or Boyd and Richerson models achieve regularization through a bias in mean behavior at the population level, actual human regularization behavior occurs via changes to the sampling error in the population's responses. Participant behavior also departs from the basic model in the 5:5 condition. Here, frequency-dependent selection models would predict no change: the variants do not differ in terms of relative frequency and therefore can not be differentially copied on the basis of their frequency. However, many participants regularized a 5:5 observation ratio by overproducing one of the variants, seemingly at random (refer back to Figure 3.12). This means that the regularization behavior can not be modeled in terms of a basic frequency-dependent selection or copying model in which no bias acts to change or distort the relative frequency of variants when they occur in equal proportions, because human regularization biases certainly do change the relative frequency of variants when they occur in equal proportions.

Chapter 7

Conclusions about inductive evolution and regularization

If you are arriving here from the previous chapters, and *especially* the previous one, this conclusion will hopefully explain what just happened to you and tell you that everything is going to be ok. And to the dear and mysterious reader, who is not one of my two supervisors or one of my two examiners, I have fashioned the conclusion to serve as a stand-alone recapitulation of this anthropologist-born-again-cognitive-scientist's take on cultural evolution. This thesis, as many theses are, is a roadmap of a research agenda. So now I will tell you where we have been, what places are worth returning to with the family (I mean your lab), and where I think we should go from here.

The main point of this thesis is that cultural evolution is unique. It encompasses a range of evolutionary dynamics that can be adequately-described by different models of biological evolution, but it also encompasses processes that have no clear analogue in biological evolution. All research that focuses on the forces of evolutionary change that are unique to culture, inevitably adds to the general theory on evolution and broadens scientific knowledge of the possible forms of evolution that exist.

Through five experiments, I targeted the methodological interface of population genetics and experimental psychology to understand how cognitive biases shape the evolution of culturally transmitted data sets. To cut close to this interface, I investigated inductive biases on frequency learning and production to understand why language, humanity's premier product of cultural transmission, contains so little unpredictable variation. Language learners possess a regularization bias, which leads them to eliminate variation in language, and a variety of cognitive mechanisms feed into this bias.

Experiment 1 established a baseline for unbiased frequency learning behavior and in doing so, identified probability matching as a cognitive basis for cultural drift. The result of this experiment showed that neutral copying obtained from an inductive process can be much more accurate (in the sense of preserving the relative frequencies of variants in the population) than a genetic process of randomly sampling alleles from a population. This is because an inductive process is affected by the *amount* of data it sees, whereas a genetic process is only affected by the *proportion* of alleles in the previous generation. If a person views one red and two blue marbles being drawn from a bag, they may be less certain of the proportion of marbles in the bag than if they had seen 100 red and 200 blue marbles drawn. If you ask a person to tell you 10 more draws that are likely to come from that bag, then you will get different kinds of responses depending on how much data you let the person see. However, if a random mating process samples 10 alleles from a population of 1:2 or 100:200 allele types, the results will be identical. Therefore, inductive evolution defines a type of cultural drift which can neutrally maintain variation in the population longer than models of genetic drift can, especially if large data sets are seen. This suggests that importing population genetics models of sampling processes as a baseline for cultural evolution is inappropriate. Instead, better baselines can be informed by understanding the sampling error associated with a wide variety of cognitive models and models of inductive inference.

Experiment 2 built upon the basic frequency learning design of Experiment 1 to engage different sources of regularization biases in adult learners. Domain-general drivers of regularization were investigated by having participants learn about marbles being drawn from one container, or marbles being drawn from several containers concurrently. In the one-item task participants probability matched, as in Experiment 1, whereas in the six-item task they regularized (i.e. they tended to indicate that only the more common marble color would come out of each container). This behavior may be rooted in memory constraints, such as forgetting about low-frequency variants when cognitive load is high. Language learners are constantly tracking the statistics of linguistic variants at all levels of language (its phonology, lexicon, and syntactic structures) and this kind of domain-general constraint on frequency learning may be part of the reason why language learners regularize. However, language learners also regularize when they are only learning about one linguistic item at a time. This domain-specific driver of regularization was investigated by reframing the basic frequency learning task from Experiment 1 in the linguistic domain. Although concurrent frequency learning and linguistic frequency learning both elicit regularization behavior, nei-

ther one on its own accounts for the full amount of regularization behavior obtained when participants learn about multiple linguistic frequencies concurrently. Therefore, both domain-general and domain-specific aspects of language learning have a role in explaining regularization in language. Cultural data sets are shaped as they pass through cognition, which is an integrated system forged by our genes, individual experience with the world, and social experience with others. All three of these sources will contribute to the overall form that any product of cultural transmission takes and thus, none of these on their own can predict the evolutionary trajectories that something such as language will follow. Any attempt to understand the cultural evolution of behavior by restricting our explanation to the subset of cognition that processes socially-acquired information will necessarily be an incomplete story.

Experiment 3 and 4 took a closer look at memory constraints involved in concurrent frequency learning. These experiments showed that the level of regularization behavior that participants produced was robust, across training sets that differed markedly in the amount of variation they contained, and across different training and testing regimes that modulated the cognitive load of the task. Although these manipulations did affect the kind of frequency information that participants encoded, regularization behavior seems to be heavily modulated by what participants decide to *do* with this frequency information when it comes to the production of linguistic utterances. This raises the point that the results of artificial language learning experiments may be sensitive to aspects of task framing, pragmatic factors, and participants' perceived goal of the experiment.

Experiment 5 addressed the pragmatic factors underlying linguistic regularization by investigating how two individuals utilize shared frequency information to coordinate in a non-linguistic task. This functional use of frequency information also led to regularization behavior. In this experiment, a pair of participants each drew several marbles from the same bag, but they could not see what their partner had drawn. Then in an implicit coordination game, they were both asked to write down the same color without communicating with one another. When participants made draws of five blue marbles and five red marbles, they achieved coordination on the blue marble better than chance. However, when participants drew three of one color and seven of the other color, they ignored the color identity of the marbles and successfully coordinated on the one they had observed in the higher frequency. This implicit coordination game was repeated ten times *without feedback*. In both conditions, many participants maximized by consistently choosing one color in each of the ten coordination games. Participant behavior in this task led to the same amount of regularization as in Experiment 2, when par-

ticipants had to name one object ten times in the absence of an explicit pressure to coordinate with a partner. This suggests that artificial language learning tasks may elicit coordination-based regularization strategies from learners as they try to provide the "best" answers in test trials. If this is a maximization strategy, it will follow a simple rule: "always choose the most frequent variant", and therefore ignore the nuanced differences in the statistical information that participants have encoded.

Several important methodological points were raised in this thesis. First, I proposed an information-theoretic definition of regularization behavior as a drop in the conditional entropy of linguistic variants on contexts (Section 3.1). Although this is an obvious measurement of linguistic structure such as words conditioned on meanings, or determiners conditioned on noun classes, it can also be seamlessly used to quantify the variation in a distribution of linguistic variants irrespective of their contexts. Most linguistic regularization experiments collect data of this latter type and discuss distributional variation in terms of "scatter" and "learning context variability" and report statistics for the highest-frequency variant. All distributions over linguistic variants can be quantified in terms of entropy, which could be a powerful universal measure from comparing the amount of regularization elicited in different experiments, and across different linguistic units at all levels of language (such as phonological forms, the lexicon, morphosyntactic markers, syntactic structures, and word orders).

Second, in Chapter 5, an analysis of some Bayesian models of inductive inference showed that two types of beta-binomial models (the averager and maximizer) are equivalent to the Wright-Fisher model of genetic drift, without mutation, when the averager's inductive bias favors complete regularity, and when the maximizer's inductive bias is *un*biased. Because the Wright-Fisher model of genetic drift defines neutral evolution, the cognitive equivalent, intuitively, should occur in the absence of an inductive bias. Because only the maximizer produces unbiased behavior for an unbiased prior, the maximizer seems to be the appropriate model for relating inductive biases to behavioral biases, especially when those behaviors evolve over time. Reali and Griffiths (2010) have shown that the averager model is equivalent to the Wright-Fisher model of genetic drift *with mutation* for all possible values of its bias (because every bias value can be matched with a mutational value). This result implies that cognitive biases are equivalent to mutational pressures on culturally evolving data sets. This makes sense for the averager model, because all values of its inductive bias (besides maximum regularity) lead to behavior that is biased toward variability. But it makes less sense for the maximizer model, which produces both variability-biased behavior (i.e.

anti-conformity copying) and regularity-biased behavior (i.e. conformity-copying) because conformity-copying mutation is an odd concept (though it may be a new concept unique to inductive evolution). Since the maximizer is the only one of these three models that can actively eliminate variation from a pool of cultural variation (via conformity copying), the maximizer model may be a more promising avenue for future equivalence results, such as those for Wright-Fisher models with selection. Finally, we saw that neither of these three Bayesian models of frequency learning were capable of regularizing in the way that human learners do. Despite this shortcoming, all of these models are still models of inductive evolution and therefore, do constitute theoretically possible models of cultural evolution. With regard to unique aspects of cultural evolution, the sampler model reiterates a point made in Experiment 1, which demonstrated that cultural evolution can follow non-binomial sampling processes. Whereas probability matching defines cultural drift with *lower* sampling error than binomial drift, the sampler model defines cultural drift with *higher* sampling error than binomial drift. This would lead to faster elimination of neutral variation than models of genetic drift would predict and demonstrates yet another way in which cultural evolution breaks the binomial sampling assumption so common in descriptions of genetic evolution.

Third, in Chapter 6, I addressed the problem of linkage with estimates of empirical transition matrices derived from the experimental data gathered in Experiments 1 and 2. From these estimated transition matrices, the outcome of cultural evolution can be estimated in terms of the stationary distribution over all behaviors after an infinite number of generations of learners. Here, we saw that the behavior of a single generation of learners does not necessarily provide a good estimate of the ultimate form, or typological distribution, that culturally transmitted behavior takes. This is because culturally transmitted behaviors move through a complex space of transition probabilities and experimentally determining what one slice of this space looks like can be undermined by unusually high-probability transitions via states which have not been investigated. Therefore, we should not, by default, expect that the distribution over individual behaviors in one generation will carry complete information about the distribution over behaviors after cultural transmission has taken place. Certainly, this end-state distribution is completely determined by individuals' biases and the behavior they produce, but the *pathways* that form when behaviors are linked up via cultural transmission are the ultimate determiner of typological distribution.

Inductive evolution is the unification of cultural evolution and cognitive science. It defines a unique force of evolution that operates in cultural transmission systems and is necessarily cognitive. Behaviors are never directly copied, but

always reverse engineered in a cycle of perception, processing, and production. Higher-level descriptions of behaviors in terms of direct copying mechanisms are indispensable for generating and culling hypotheses about the forces of evolution at play in any particular cultural data set, such as Paleo-Indian projectile points, West German pottery shards, first names in the U.S. census, and lexical items in particular genres of literature. However, macro-level analyses of cultural change will always underspecify the mechanism because macro-level patterns are multiply realizable. Because the locus of cultural change is cognition, this is the most informative level for studying the forces of cultural evolution.

Appendix A

Experiment instructions

A.1 Experiment 1

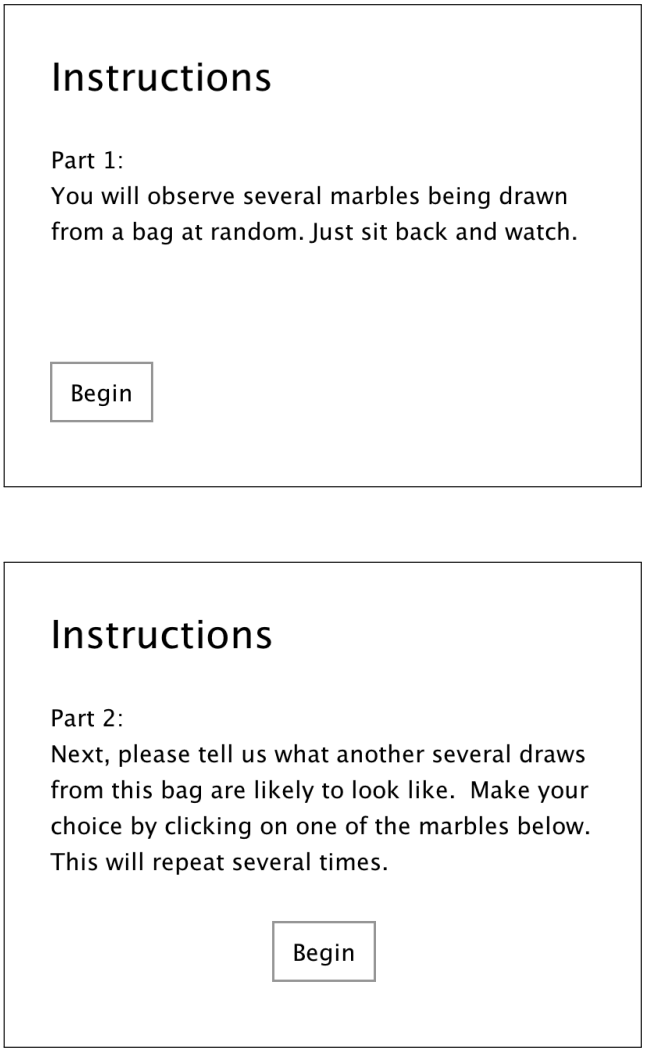


Figure A.1: Screen shots of the observation phase instructions (top) and the production phase instructions (bottom).

A.2 Experiment 2

	marbles1	*words1*
Part 1	You will observe several marbles being drawn from a bag at random. Just sit back and watch.	You will observe one object being named several times. Just sit back and watch.
	(10 observation trials)	(10 observation trials)
Part 2	Next, please tell us what another several draws from this bag are likely to look like. Make your choice by clicking on one of the marbles below. This will repeat several times.	Next, you will be shown the same object again. Please name this object like you saw in Part 1. This will repeat several times.
	(10 production trials)	(10 production trials)
Part 3	The container you saw had 100 marbles inside. This container could have one or two different colors of marbles in it. For this container, please tell us how many marbles of each color were in it. Please make your best guess.	The words you saw come from another language than English. The object you saw can be named with one or two words in this language. For this object, please tell us what percent of the time it is named with each word in this language. Please make your best guess.
Part 4	In Part 1, you saw 10 marble draws. How many times did you see each color? Please make your best guess. In Part 2, you made 10 marble draws. How many times did you choose each color? Please make you best guess.	In Part 1, you saw the object being named 10 times. How many times did you see each word? Please make your best guess. In Part 2, you named the object 10 times. How many times did you choose each word? Please make your best guess.
Part 5	(Exit questionnaire)	(Exit questionnaire)
Part 6	Thank you for participating. Please write down this completion code and enter it on the HIT page so that we know you have completed the task.	(same)

	marbles6	*words6*
Part 1	You will see 6 different containers. Each container is filled with lots of marbles. We will randomly draw marbles out of each container and show them to you. Try to get a feel for what the draws from each container look like. This part takes about 3 minutes. Please pay attention.	You will observe six different objects being named several times. Just sit back and watch. This part takes about 3 minutes. Please pay attention.
	(60 observation trials)	(60 observation trials)
Part 2	Next, please tell us what some more random draws from the containers are likely to look like. Each time we show you a container, make your choice by clicking on one of the marbles below. This will repeat several times. This part takes about 3 minutes.	Next, you will be shown the same objects again. Please name the objects like you saw in Part 1. This part takes about 3 minutes.
	(60 production trials)	(60 production trials)
Part 3	Each container had 100 marbles inside. And each container could have one or two different colors of marbles in it. For each container, please tell us how many marbles of each color were in it. Please make your best guess.	The words you saw come from another language than English. Each object you saw can be named with one or two words in this language. For each object, please tell us what percent of the time it is named with each word in this language. Please make your best guess.
Part 4	(none)	(none)
Part 5	(Exit questionnaire)	(Exit questionnaire)
Part 6	Thank you for participating. Please write down this completion code and enter it on the HIT page so that we know you have completed the task.	(same)

Part 3 instruction screen shots

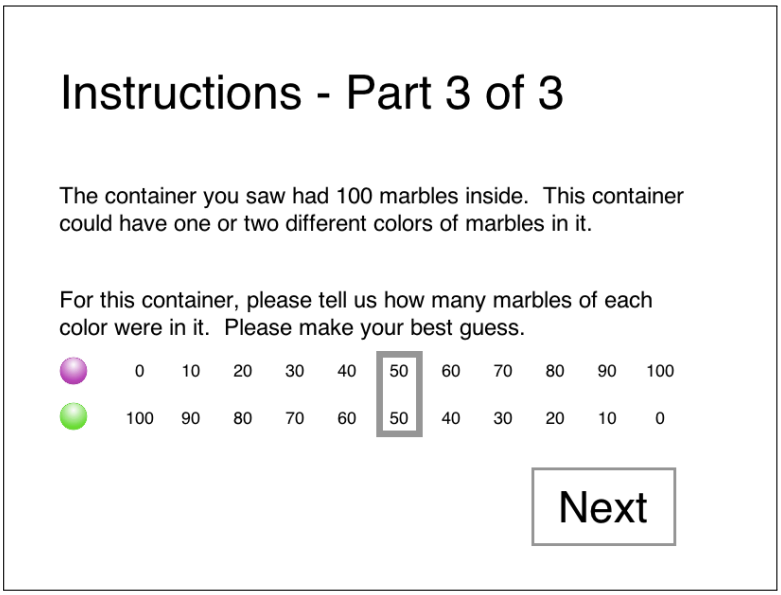


Figure A.2: Part 3 instruction screen shot for *marbles1* with answers selected.

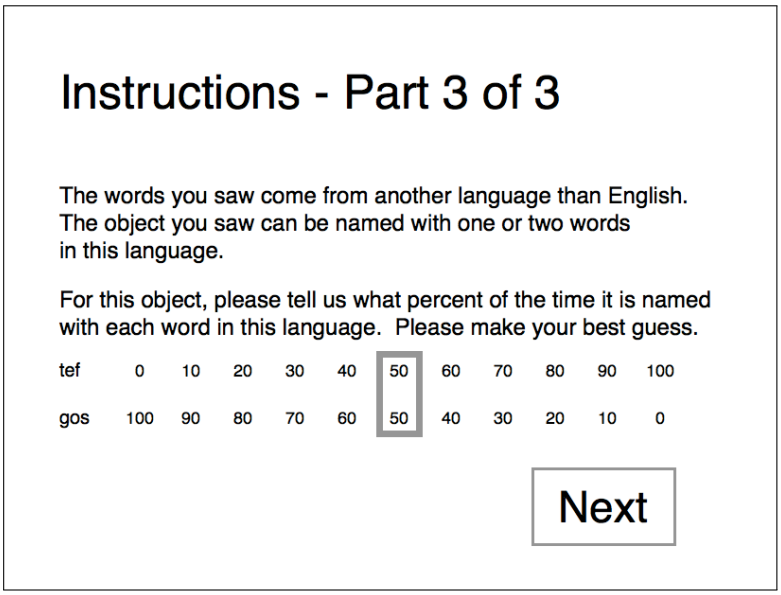


Figure A.3: Part 3 instruction screen shot for *words1* with answers selected.

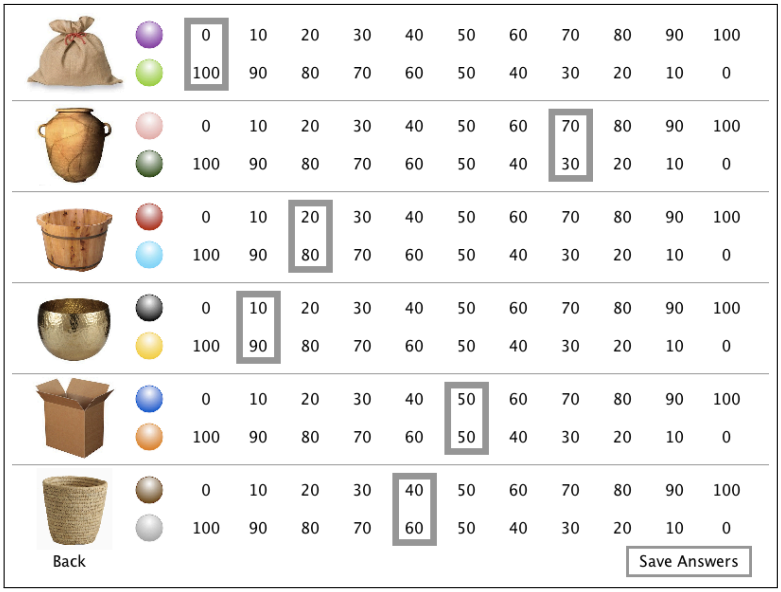


Figure A.4: Part 3 instruction screen shot for *marbles6* with answers selected.

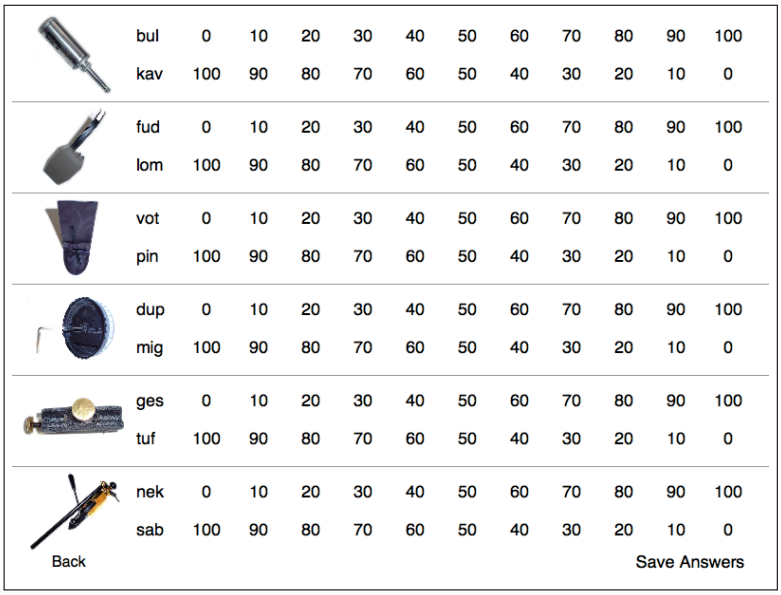


Figure A.5: Part 3 instruction screen shot for *words6* with no selections.

Part 4 instruction screen shots

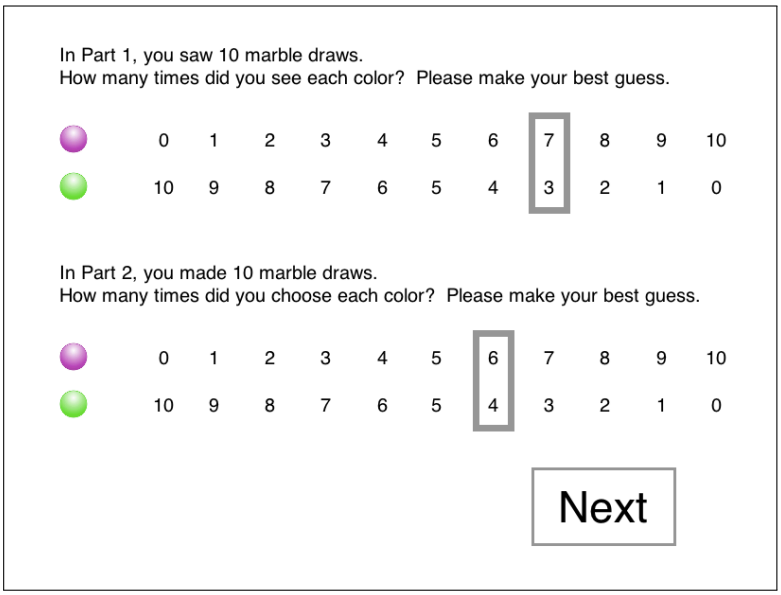


Figure A.6: Part 4 instruction screen shot for *marbles1* with answers selected.

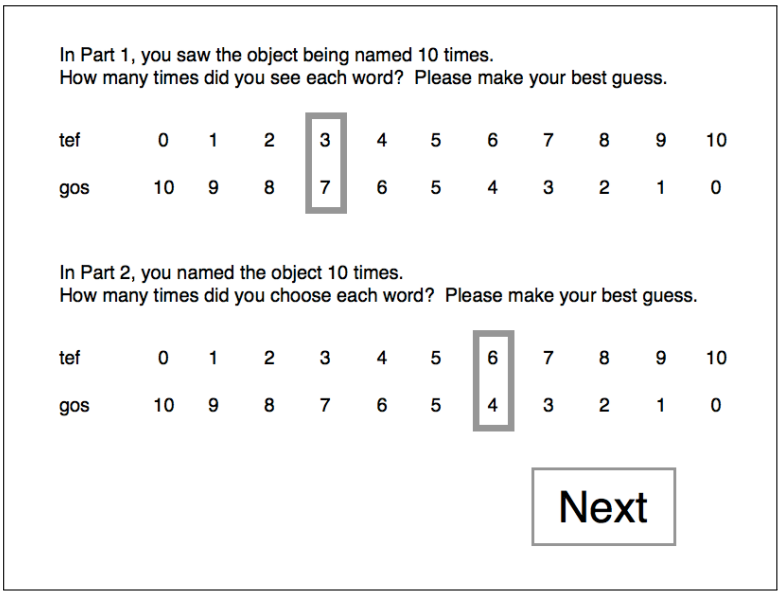


Figure A.7: Part 4 instruction screen shot for *words1* with answers selected.

A.2.1 Exit questionnaires

Participants answered each question by selecting "Yes" or "No".

marbles1 and *marbles6*

People complete this task in different ways, please tell us what you did:
1) I used a pen or pencil during the task.
2) I tried to count the number of marbles of each color.
3) I thought the marble draws were random.
4) I thought the marble draws followed a predictable sequence.
5) I tried to memorize the exact sequence of marbles I saw.
6) In Part 2, I tried to produce the same sequence of marbles I saw.
7) In Part 2, I tried to produce a random sequence of marble draws.

words1 and *words6*

People complete this task in different ways, please tell us what you did:
1) I used a pen or pencil during the task.
2) I tried to count the number of times I saw each word.
3) I thought the two words appeared randomly.
4) I thought the words appeared in a predictable sequence.
5) I tried to memorize the exact sequence of words I saw.
6) In Part 2, I tried to name the object with just one of the words.
7) In Part 2, I tried to name the object with both words.

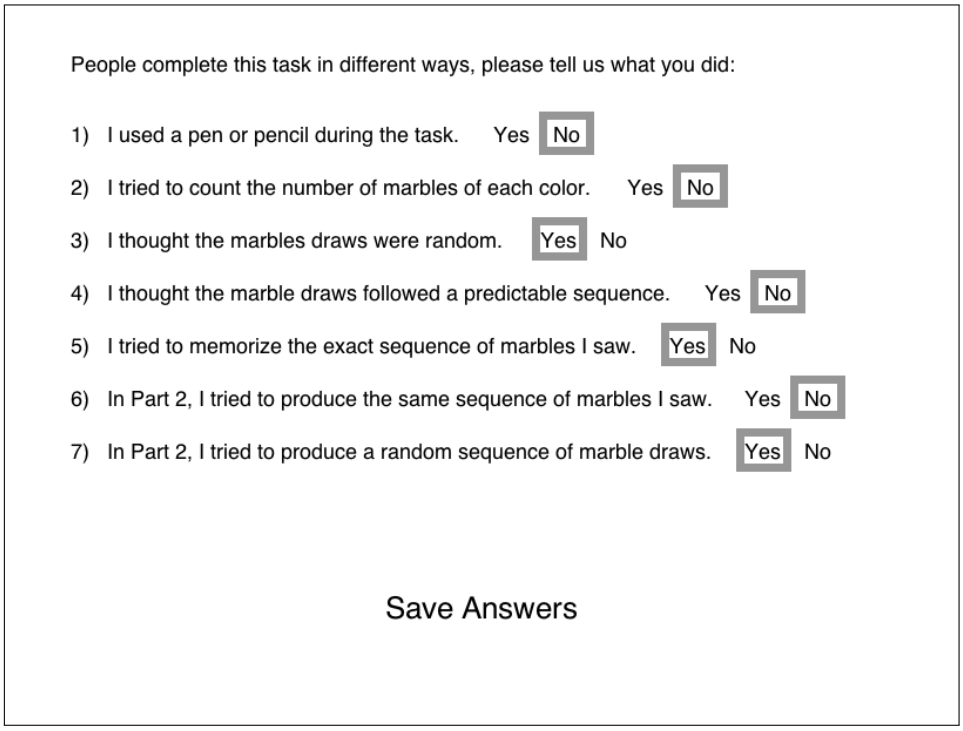


Figure A.8: A screen shot of the exit questionnaire with answers selected.

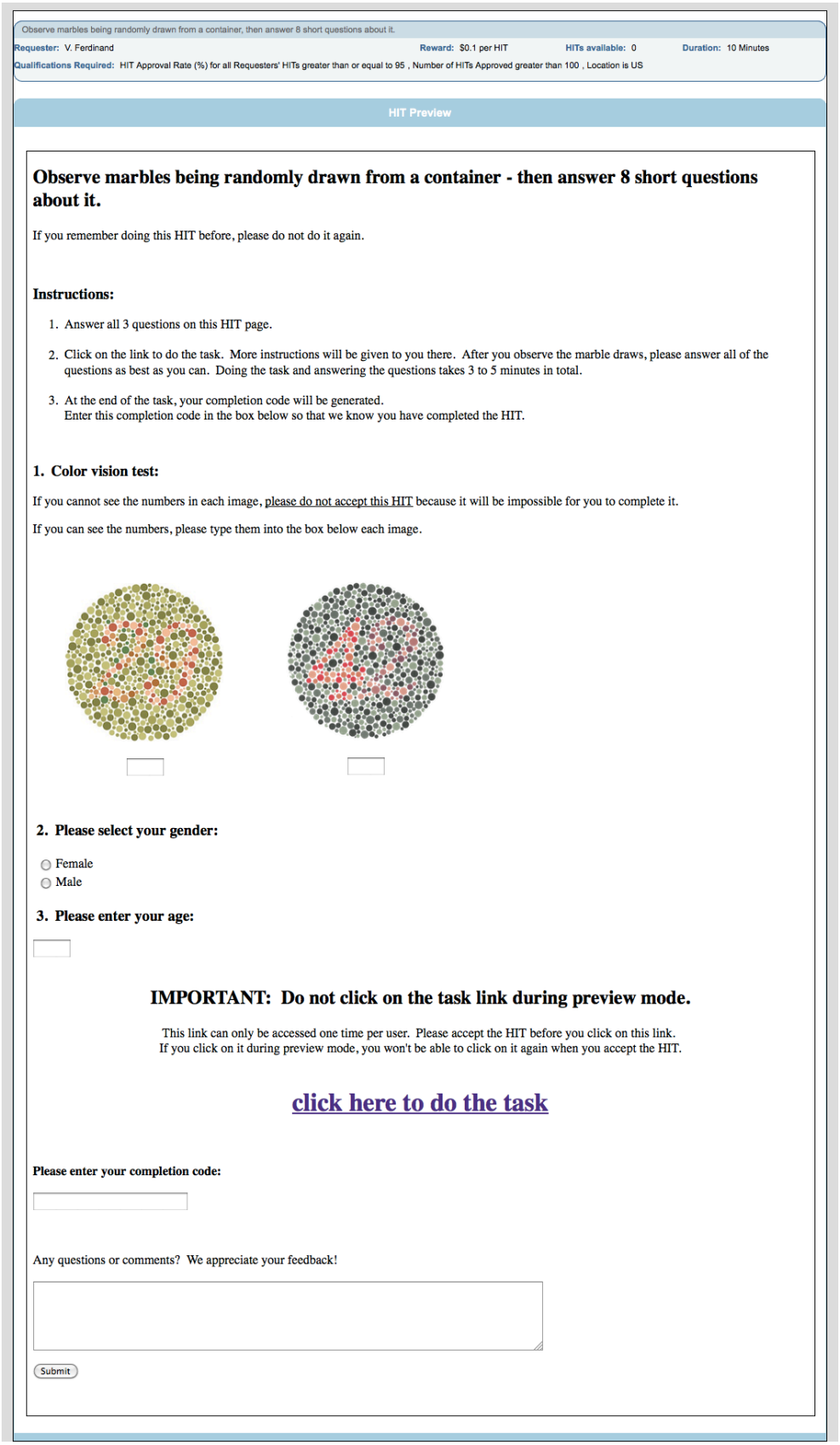


Figure A.9: The MTurk HIT page as seen by participants in condition *marbles1*.

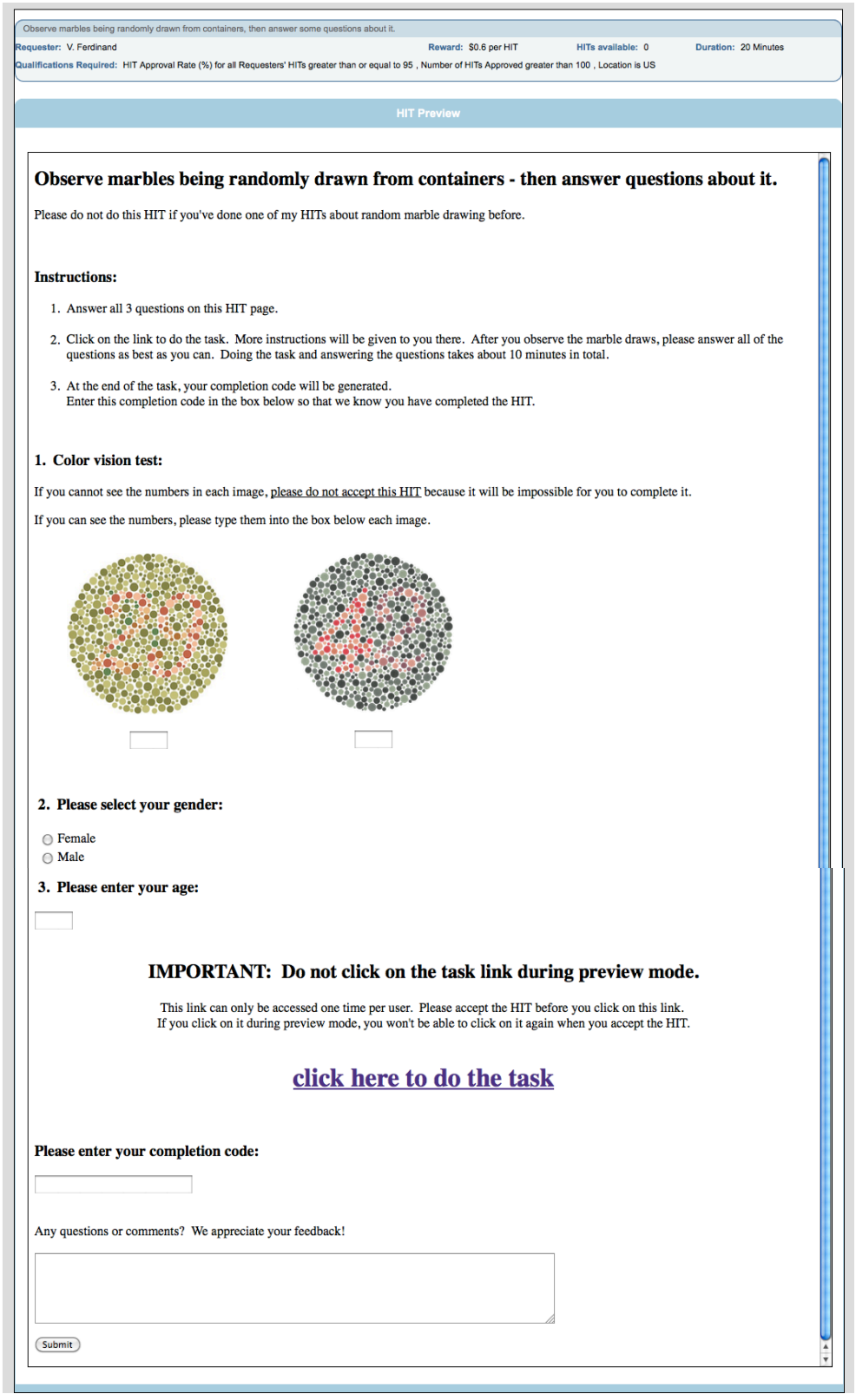


Figure A.10: The MTurk HIT page as seen by participants in condition *marbles6*.

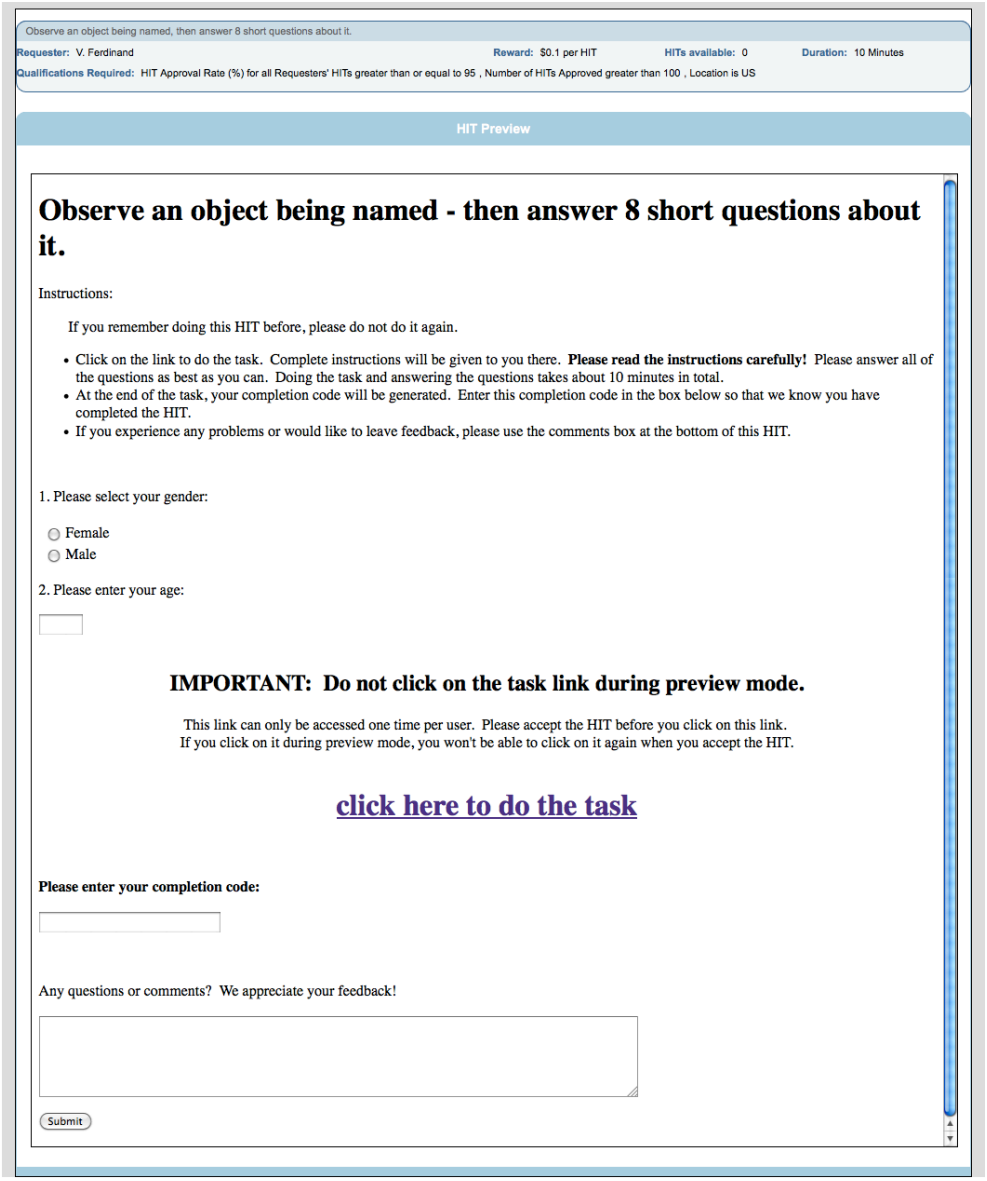


Figure A.11: The MTurk HIT page as seen by participants in condition *words1*.

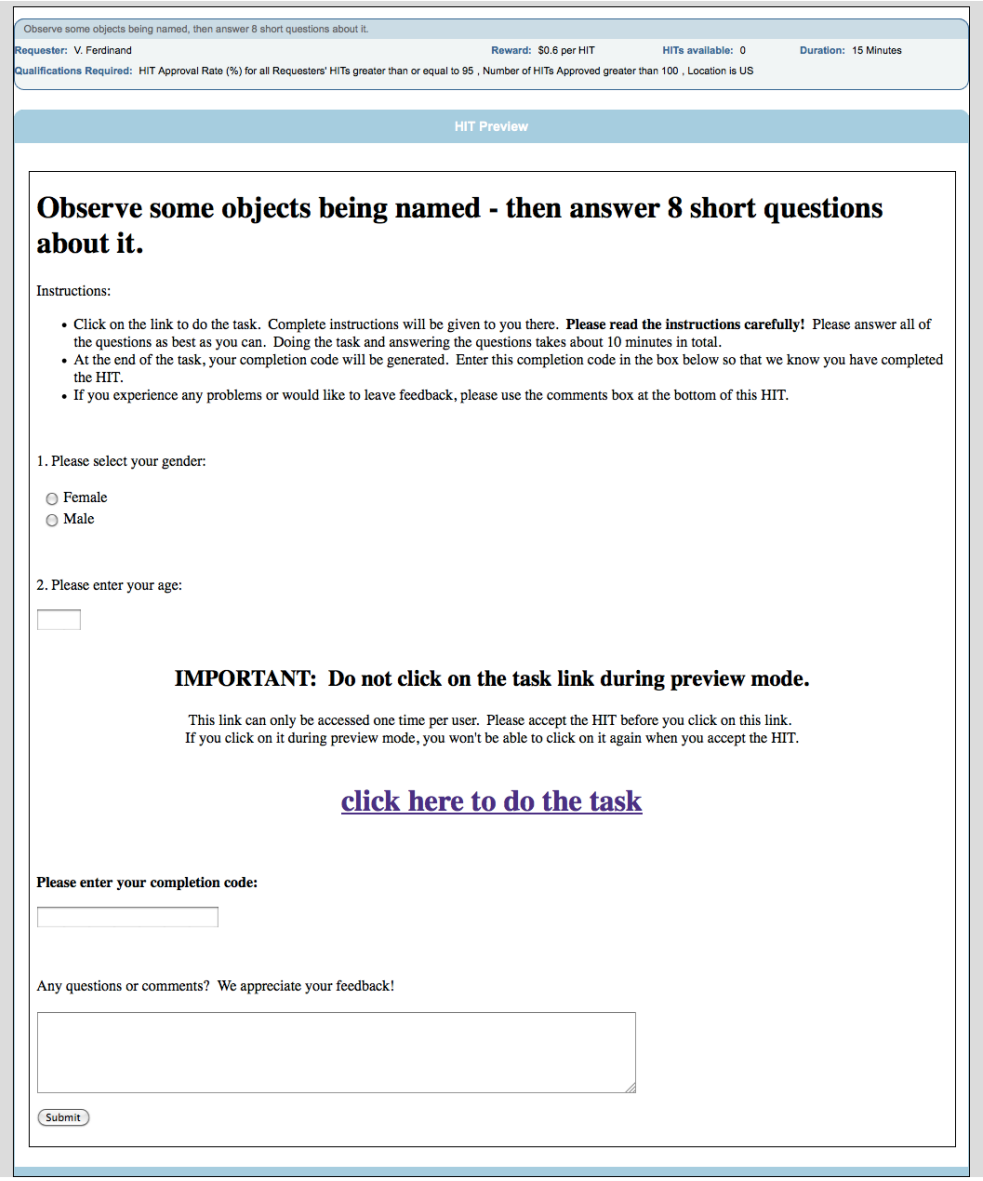


Figure A.12: The MTurk HIT page as seen by participants in condition *words6*.

A.3 Experiment 5

A.3.1 Verbal instructions

Part 1

Welcome, I'm Caroline and this is Vanessa, and thank you for participating in our research. Before we begin, we would like to ask you to read through this consent form and, if you accept, please sign it.
[CONSENT FORM]

Then, before we start, we would like to do a little color vision test, since the experiment involves differently colored marbles, and we want to make sure that you can distinguish these.
[VISION TEST]

The experiment consists of two parts. We will now give you the instructions for part one. Please take your time to read through them and make sure that you understand them.
[WRITTEN INSTRUCTIONS]

As you have read, for this experiment you will be working together as a pair. You will both draw a set of marbles from a bag. We will alternate between you two to draw marbles. So first you draw one, look at it, put it back, then you draw one, look at it, put it back, and then you again and so on.

When doing this experiment, we would normally use a soundproof booth, however since we do not have the means for that we would like to ask you to stay silent during the experiment. It is important that you do not engage in verbal or other communication with your partner. To this end, we will position you with your backs to each other, so that you cannot see your partner, or their marble draws. [REPOSITION CHAIRS]

I will hold the bag, and Vanessa will record the sequences you draw from it. If you have any questions, please ask them now.

Part 2

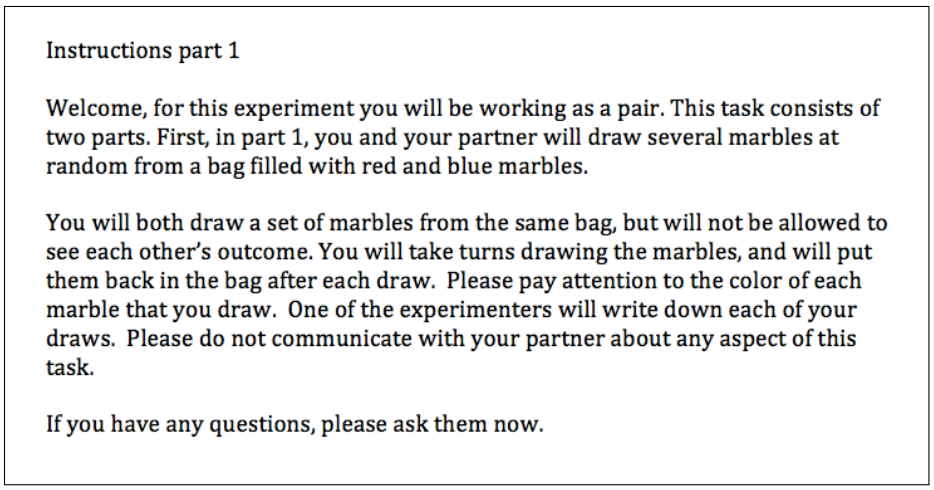
Now we will give both of you a card. On this card please write down a marble color. *The only goal in this part of the experiment is to write down the same color on the paper as your partner does on their paper.* If you succeed, you will receive a small reward.

Good. Now we will give both of you another card. On this card please write down a marble color. *The only goal in this part of the experiment is to write down the same color on the paper as your partner does on their paper.* If you succeed, you will receive a small reward.

[repeat instruction verbatim until 10 responses are collected]

Figure A.13: Complete verbal instructions for Experiment 5. The experimenter had this script on her clipboard and delivered it identically to all participants.

A.3.2 Written instructions


Instructions part 1

Welcome, for this experiment you will be working as a pair. This task consists of two parts. First, in part 1, you and your partner will draw several marbles at random from a bag filled with red and blue marbles.

You will both draw a set of marbles from the same bag, but will not be allowed to see each other's outcome. You will take turns drawing the marbles, and will put them back in the bag after each draw. Please pay attention to the color of each marble that you draw. One of the experimenters will write down each of your draws. Please do not communicate with your partner about any aspect of this task.

If you have any questions, please ask them now.

Figure A.14: The written instructions for Experiment 5. A sheet with these instructions was handed to each participant to read at the "written instructions" cue in the verbal script (shown above in A.3.1).

A.3.3 Informed consent form

Participant Identification Number:

Informed Consent Form

Name of researchers: Vanessa Ferdinand and Caroline Kamps

Your participation is valuable to us and very much appreciated. However, your participation is entirely voluntary.

Please check box

1. I agree to take part in the study carried out by Vanessa Ferdinand and Caroline Kamps. ☐

2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, including after consenting to participate. ☐

3. I understand that any information given by me may be used in future reports, articles or presentations by the research team. ☐

4. I understand that my participation and the data that I provide with my responses will be kept confidential. My name will not be recorded anywhere in the database or appear in any reports, articles or presentations. I understand that individual participants will not be identified (or will be identified only by codes). ☐

5. I have been given a copy of this consent form ☐

Signature of the participant
I understand what is involved in this research and I agree to participate in the study.

Name of Participant	Date	Signature

Signature of the researcher
I believe the participant is giving informed consent to participate in this study

Researcher	Date	Signature

Figure A.15: A copy of the informed consent form signed by participants in Experiment 5.

Appendix B

Publications

Regularization behavior in a non-linguistic domain

Vanessa Ferdinand (v.a.ferdinand@sms.ed.ac.uk), Bill Thompson (bill@ling.ed.ac.uk),
Simon Kirby (simon@ling.ed.ac.uk), Kenny Smith (kenny@ling.ed.ac.uk)
Language Evolution and Computation Research Unit
School of Philosophy, Psychology & Language Sciences, University of Edinburgh
Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK

Abstract

Language learners tend to regularize unpredictable variation and some claim that is due to a language-specific regularization bias. We investigate the role of task difficulty on regularization behavior in a non-linguistic frequency learning task and show that adults regularize variable input when tracking multiple frequencies concurrently, but reliably reproduce the variation they have observed when tracking one frequency. These results suggest that regularization behavior may be due to domain-general factors, such as memory limitations.

Keywords: frequency learning; regularization; probability matching; Bayesian models;

Introduction

Languages contain very little unpredictable variation (Chambers et al., 2003) and language learners tend to regularize the inconsistent input they encounter (Reali & Griffiths, 2009; Hudson Kam & Newport, 2009, Smith & Wonnacott, 2010). For example, English contains two forms of the indefinite article *a* and *an*, but a deterministic rule (based on the initial phoneme of the following noun) governs the use of these two variants. Why are languages regular, and what drives learners to eliminate free variation in language? Some have suggested that we come to the task of language learning with the expectation that languages are regular and that this expectation takes the form of a language-specific innate bias (Bickerton, 1984; DeGraaff, 1999; Lumsden, 1999; Becker & Veenstra, 2003). Others claim that linguistic regularization can be explained by domain-general learning mechanisms, such as the effects of memory limitations on the type of variation that learners produce (Hudson Kam & Newport, 2005, 2009; Hudson Kam & Chang, 2009). Hudson Kam and Newport (2005, 2009) have shown that children tend to regularize free variation, whereas adults maintain it by probability matching, and attribute this difference to children having lower working memory capacity than adults. Newport (1990) demonstrated that children have more of a limited ability to learn from inconsistent input and Hudson Kam and Chang (2009) showed that adults probability matched more when word retrieval was made easier and regularized more when it was difficult, further corroborating their claim that memory limitations can lead to regularization, although see Perfors (2012) for an account of restricted memory encoding that does not lead to regularization.

A similar effect of memory limitations can be found in a non-linguistic tasks. In a study with adults, Kareev et al. (1997) reported an effect of individual differences in working memory capacity (as determined by a digit-span test) on participants' perception of the correlation of two probabilistic variables. Participants with lower capacity overproduced the most common variant, whereas participants with higher capacity did not. Regularization is also modulated by the number of variables in a task; adults regularized slightly more when predicting which of three lights will flash next than when predicting for two lights (Gardner, 1957).

In this paper, we explore the effect of tracking single versus multiple frequencies on the regularization behavior of adults in a non-linguistic task. We show that participants probability match when tracking a single frequency, but regularize when tracking six frequencies concurrently. Because concurrent frequency learning is a prominent aspect of language learning (Saffran, Alin & Newport, 1996), and also elicits regularization in a non-linguistic task, this is consistent with a domain-general account of the observed regularization bias in language, possibly attributable to limited working memory.

Frequency learning experiment

Participants 381 participants were recruited via Amazon's Mechanical Turk crowdsourcing platform and completed our experiment online. 37 participants were excluded on the basis of the following criteria: failing a color vision test (2), self-reporting the use of a pen or pencil during the task (14), not reporting their sex or age (2), or having previously participated in any of our experiments, as determined by their user ID with MTurk (19). More participants were recruited than necessary with the expectation that many would be excluded by these criteria. Once the predetermined number of participants per condition was met, data from the last participants was excluded, totaling 24 participants across all conditions and tasks. All excluded participants received the full monetary reward for the task. The average monetary reward per participant, converted to an hourly rate, was $2.64. Of the final 320 participants, 184 are female, and the mean age is 36 (min = 18, max = 69), with a standard deviation of 12 years.

Materials The experiment was coded up as a java applet that ran in the participant's web browser in a 600x800-pixel field. Photographs of 6 different containers (a box, pouch, jar, bowl, bucket, and basket) and graphically generated images of marbles in 12 different colors (blue, orange, brown, grey, black, yellow, red, teal, olive, pink, purple, and lime) served as stimuli.

One-item task This experiment consisted of a training phase in which participants observed a series of 10 marble draws from a bag, and a testing phase in which participants were asked to produce another several likely draws from the

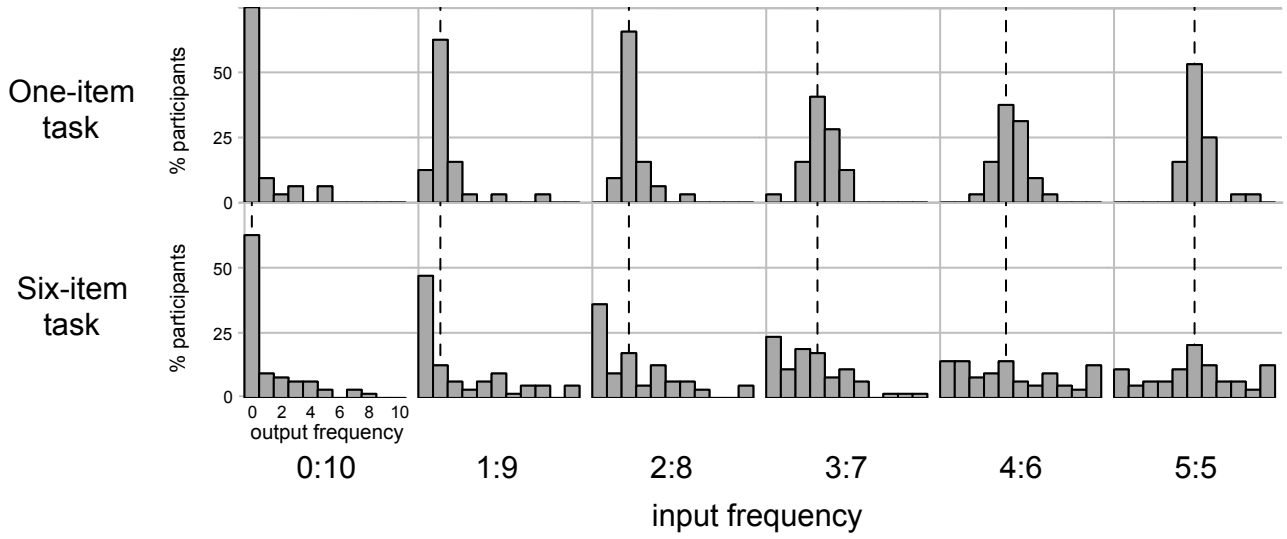


Figure 1: Each pane displays the percentage of participants that responded with a given output frequency of the minority marble (*m*) during testing. Columns are the input ratio of *m*:*M* during training. Dashed lines mark the input frequency of *m*. In the one-item task, participants probability matched, reproducing the input ratio with high fidelity. This task was between-subjects; each participant was trained on one input ratio only. In the six-item task, participants were more likely to regularize than to reproduce the input ratio. This task was within-subjects; each participant was trained on all six input ratios concurrently.

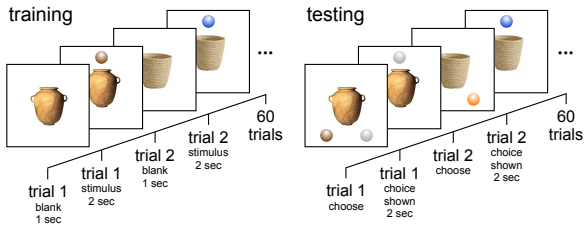


Figure 2: Training and testing trials for the six-item task.

same bag. In each training trial, a picture of the bag was displayed for 1000 milliseconds and then a marble (blue or orange) appeared over the bag for 2000 milliseconds. There were 10 training trials, with no break between trials. In each testing trial, the bag was displayed with the two marble colors below. Participants mouse clicked on a marble to make their choice of one draw from the bag. Their choice was displayed above the bag for 2000 milliseconds and then the next testing trial began. There were 10 testing trials with no breaks between trials. Locations (left or right) of the blue and orange marbles were held constant across test trials for each participant, but counterbalanced across participants.

A fixed ratio of blue to orange marbles was shown in the training phase. Each participant was randomly assigned to one of 6 training conditions based on this ratio. The color of the training ratio's minority marble (*m*) and majority marble (*M*) was counterbalanced across participants. All possible ratios of *m*:*M* were tested and will be referred to as the 0:10, 1:9, 2:8, 3:7, 4:6, and 5:5 conditions. 192 participants took part in this task, with 32 in each condition.

Six-item task This task is based on the word frequency learning task from Reali and Griffiths (2009). Participants observed 10 marble draws each from six different containers, totaling 60 marble draws (see Figure 2). Each container was associated with 2 unique marble colors (12 unique marble colors were therefore used). Training and testing trials were identical to the one-item task. Each container was uniquely associated with one of the possible ratios specified by condition 0:10, 1:9, 2:8, 3:7, 4:6, and 5:5 above. Thus, the six-item task is a within-subject version of the one-item task, with the addition that training and testing trials from all six conditions are interleaved. Assignments of a ratio and marble colors (in predefined color pairs) to each container was randomized per participant. 64 participants took part in this task. Two additional versions of this experiment were also run; one where all 6 bags were in condition 0:10 (each container was mapped to one color only) and one where all 6 containers were in condition 5:5. Each of these versions was completed by 32 new participants.

Experiment results

Participants in the six-item task were more likely to regularize their responses per container than participants in the one-item task. Here, we refer to regularization as the production of a more extreme ratio than that observed during training,

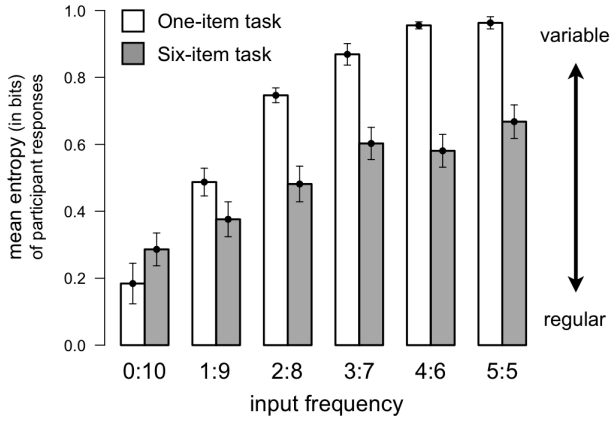


Figure 3: Difference in mean entropy scores between tasks, for each input ratio. Each participant's sequence of marble draws during testing was converted into an entropy score. Lower scores denote greater regularity within a response. Participant responses were significantly more regular in the six-item task than in the one-item task for input ratios 3:7, 4:6, and 5:5. Error bars show the standard error of the mean.

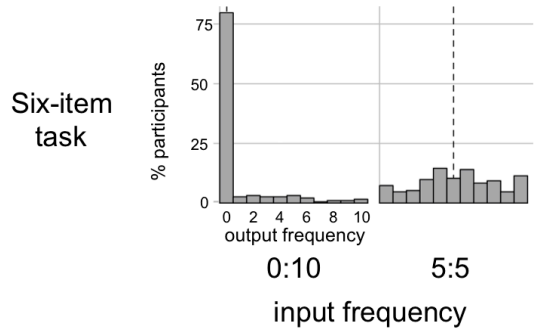


Figure 4: Distribution of participant responses for two additional versions of the six-item task, where all items contained the same input ratio of m:M. One group of participants was trained on all 0:10 ratios and another group was trained on all 5:5 ratios.

where 0:10 and 10:0 are the most extreme ratios and 5:5 is the least extreme. The distributions of participant responses are shown in Figure 1. Each pane displays the percentage of participants that responded with a given output frequency of m, per input frequency and per task. In the one-item task, participants probability matched; the mode of the population is on the input frequency of m, meaning that the most common response was perfect reproduction of the ratio observed during training. In the six-item task, visual inspection suggests that participants did not reproduce the training ratios with as high fidelity. Most participants regularized by overproducing the majority marble (all mass in the bars to the left of the dotted line) and a large number of responses are fully regular, meaning the output frequency of m is 0 or 10.

To better assess the different degrees of regularization between tasks, we calculated the entropy of each participant's sequence of test choices. This quantifies the amount of variation (in bits) with a value between 0 and 1; where 0 denotes a completely regular sequence (i.e. a series of all blue marble draws) and 1 denotes a maximally variable sequence (i.e. a series of 5 blue and 5 orange draws, in any order). This allows us to refine our definition of regularization as the overproduction of one marble, such that the entropy of the participant's testing choices is lower than that of their training observations. The mean entropy scores of participant responses per input frequency are shown in Figure 3.

A linear mixed effects regression analysis showed a significant effect of task on entropy scores, $t(34) = -7.226, p < .001$, and a significant effect of input frequency on entropy scores, $t(34) = -10.832, p < .001$. This means the two tasks elicited different amounts of regularity within participants' responses and that participants' responses were modulated by training frequencies; they noticed differences in the input frequencies and this affected their responses. A significant interaction of task and input frequency on entropy scores was also obtained, $t(34) = 4.570, p < .001$; participants responded differently to different input frequencies, and this pattern of responses also differed by task.

There was a significant difference in mean entropy scores between tasks for input frequencies 3:7, 4:6, and 5:5 ($W = 1427.5, p = .001; W = 1714, p < .001; W = 1585.5, p < .001$), respectively.[1] The difference in mean entropy between tasks was not significant for input frequencies 0:10, 1:9, and 2:8 ($W = 894, p = .228; W = 1184.5, p = .192; W = 1264, p = .054$[2]), respectively.

Two additional experiments were conducted to explore the possibility that regularization in the six-item task is due to interference between containers, such that ratios learned for one container get confused with ratios learned for another container. We eliminated this type of interference by training participants on 6 containers with identical ratios. Figure 4 shows participant responses when trained on all 0:10 ratios (left) and all 5:5 ratios (right). The average entropy for the all 0:10 task is significantly lower than that of the 0:10 condition in the six-item task ($W = 5061, p = .004$), but not significantly different than the 0:10 condition in the one-item task ($W = 2900.5, p = .466$). Tracking multiple 0:10 ratios is no different than tracking one 0:10 ratio, but it is different from tracking one 0:10 ratio concurrently with other ratios. This means interference may account for the errors participants make in the original six-item task when producing draws for the container they observed as 0:10. However, for the all 5:5 task, the average entropy was not significantly different from

[1] These were determined with a non-parametric t-test, the Whitney-Mann U-test, since the distributions of entropy scores are non-normal.

[2] After correction for multiple comparisons, this is not approaching significance.

the 5:5 condition in the six-item task ($W = 5892.5, p = .617$). Participants still produced 0:10 and 10:0 responses in the absence of observing these ratios during training. Therefore, interference may account for some of the differences between the one-item and six-item tasks, but this isn't the sole cause of the regularization behavior observed in the six-item task.

Frequency learning models

What cognitive processes cause regularization? So far our analyses have quantified the difference in regularity between participants' training and testing responses. In this section, we turn our focus to an internal force that can affect a learner's behavior; an inductive bias favoring certain ratios of marbles.

Bayesian model

Bayesian models provide a way to quantify inductive biases and understand their effect on behavior. We fit a beta-binomial Bayesian sampler model to participants' responses, following Reali and Griffiths (2009), and ask what prior expectation for regularity a Bayesian rational learner would need to have in order to produce the data that our participants produced.

A Bayesian rational learner uses Bayes' rule, $P(h|d) \propto P(d|h)P(h)$, to infer what proportion of marbles generated the draws that they observed. Here, each proportion is a hypothesis and the observed draws are the data. Bayes rule combines the prior probability of a hypothesis, $P(h)$, with the likelihood of the data under that hypothesis, $P(d|h)$, to arrive at a posterior probability of that hypothesis given the data, $P(h|d)$. The prior is a beta distribution over all hypotheses, Beta$(\frac{\alpha}{2}, \frac{\alpha}{2})$, where the parameter α determines whether the learner expects to see regular draws or variable draws. A learner with $\alpha < 2$ will tend to regularize their productions, a learner with $\alpha = 2$ is unbiased toward any particular proportion of draws, and a learner with $\alpha > 2$ is biased towards variability in draws. The likelihood of drawing N marbles in ratio $k : (N-k)$ from a container of marbles in proportions $p : (1-p)$ follows a binomial distribution (Equation 1).

$$P(k|p,N) = \binom{N}{k} p^k (1-p)^{N-k} \quad (1)$$

Once the posterior probability over all hypotheses has been determined, the learner must choose a hypothesis to generate testing responses from. We take the case where learners sample a hypothesis from the posterior distribution, and then sample data from this hypothesis according to its likelihood (as if the learner were randomly drawing marbles from the hypothesized proportion, with replacement, as in Equation 1).

This model defines the probability of generating all testing proportions (output states) from all training proportions (input states) and can be visualized as a transition matrix between all possible states in the system. Because our experiment covers all possible training proportions for 10 draws from a bag, we can also construct an empirical transition matrix from participant responses in each task. From here on, we switch to visualizing our data in terms of marble 1 (m_1) and marble 2 (m_2)[3]. Figure 5 (top row) shows the two empirical transition matrices and three model matrices for different values of the prior parameter α. Each value of α defines a unique transition matrix, and thus a unique pattern of behavior. For example, if a Bayesian learner observes 1 draw of m_1 and 9 of m_2, and if their prior is $\alpha = 0.01$, they are most likely to produce 0 draws of m_1 and 10 of m_2, regularizing their productions. If their prior is $\alpha = 2$, they are most likely to produce 1 draw of m_1 and 9 of m_2, probability matching their productions. And if their prior is $\alpha = 10$, they are most likely to produce 3 draws of m_1 and 7 of m_2, increasing variation in their productions. Thus, the prior used here intuitively captures a range of human behaviors in frequency learning.

The model fitting task at hand is to determine which model transition matrix most resembles the empirical transition matrix, by assigning the most likelihood to the empirical data. The prior associated with the best-fit model is the one that best explains participant behavior and gives us an idea of what biases our participants may have.

The best-fit bias in the one-item task is $\alpha = 1.55$ with a log likelihood of -413, which is equivalent to correctly predicting 20% of participant responses in this task [4]. This prior shows an expectation for a slight amount of regularity in the data set. For the six-item task, the best-fit bias is $\alpha = 1.21$ with a log likelihood of -1186, equivalent to 9% response prediction. This prior shows a stronger bias toward regularity in the six-item task than in the one-item task.

Prediction percentages are lower for the six-item task because participant responses are more variable in the this task than in the one-item task. Only deterministic processes (with one output per input) can be predicted with 100% accuracy. The ceiling on model prediction for each task was determined by fitting each data set to itself, yielding a maximum of 32% accuracy for the one-item task and 16% accuracy for the six-item task. Relative to these ceilings, the best-fit models account for 61% and 56% of participant responses in the one-item and six-item tasks, respectively.

Bootstrap model

An input-based random sampling model was also fit to the data. This model defines the transition matrix that would be obtained if participants produced their testing responses by randomly sampling 10 draws from their training observations, with replacement. In this case, each row would be a binomial where p equals the training proportion of m_1. It is important to note that this transition matrix defines the dynamics of drift in one generation and may be used as a baseline for the loss of variation that can occur in the absence of a regularization bias.

[3] marble 1 (m_1) refers to the blue marble in the one-item task, and to the blue, brown, black, red, olive, and purple marbles in the six-item task.

[4] The raw log likelihoods should not be compared between tasks, because there are a different number of observations per task. This is corrected for in the prediction percentages, which are comparable between tasks.

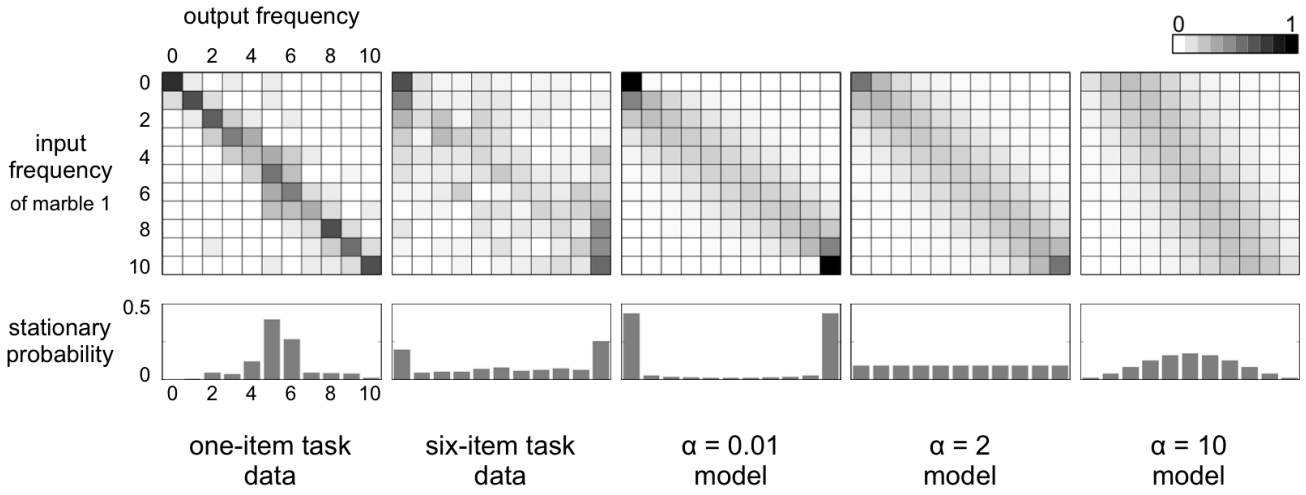


Figure 5: Transition matrices (top row) and their associated stationary distribution (bottom row) for the experimental results of the two frequency learning tasks, and for the Bayesian model showing three example bias strengths ($\alpha = 0.01, 2, 10$). Transition matrices give the probability of moving from each input frequency (the number of training trials showing marble 1) to each output frequency (the number of testing trials in which participants produced marble 1)[3]. The stationary distribution shows how often the transition matrix will produce each output frequency of marble 1.

For this model, the log likelihood of the one-item task data is -259, equivalent to 25% response prediction, and is a better fit than the best-fit beta-binomial sampler model[5]. Thus, of the models explored in this paper, drift provides the best account of our participants' probability matching behavior. However, a repeated measures Monte Carlo test shows that the standard deviation among participant output entropies in the one-item task data are significantly lower than that obtainable by drift: $p = .04$, $p = .03$, $p = .01$, $p = .003$, for conditions 2:8, 3:7, 4:6, 5:5, respectively. Although these data are well-accounted for by the drift model, they still show a quantitative difference in standard deviation, meaning that the forces behind probability matching are not truly isomorphic to drift. As for the six-item task, the log likelihood is -1076, equivalent to 6% response prediction. Here, the sampler model with a bias toward regularization is still the better fit.

Null model

This model is the transition matrix that would be obtained if participants were randomly sampling from the two testing choices each trial (i.e. not engaging in the task). Here, every row would be a binomial distribution where $p = 0.5$. For this model, the log likelihood of the one-item task data is -604, equivalent to 4% response prediction. For the six-item task, the log likelihood is -1630, equivalent to 1% response prediction. Of all models considered, this is the worst fit for both tasks, meaning that participants are not likely to be randomly sampling from their testing choices.

[5]This bootstrap model, which defines the dynamics of evolutionary drift, is equivalent to a Bayesian MAP model with $\alpha = 0$. See Reali & Griffiths (2010) for the proof.

The results of these model fits strongly suggest that participants in the six-item condition are not just performing poorly at reproducing their training proportion, but they are regularizing their responses in a way that can not be accounted for by random errors.

Learning biases and long-term behavior

In addition to comparing the transition matrices, which describe the behavior of one generation of learners, we can also look at the long-term behavior of the system, which is described by the stationary distribution of the transition matrix (Figure 5, bottom row). This distribution tells us what percent of the population we would expect to see in each state, after an arbitrarily large number of generations, if the output state of one learner served as the input state to another. Griffiths and Kalish (2007) have shown that the stationary distribution mirrors the prior distribution over hypotheses for the Bayesian sampler model utilized here. The stationary distributions of the empirical transition matrices are most interesting because these would be an estimate of our participants' regularization bias (the prior) if they were Bayesian sampler learners[6]. In line with this interpretation, the stationary distribution of the six-item task closely resembles that of its best-fit Bayesian model, which has a beta distribution Beta(0.605,0.605). However, the stationary distribution of the one-item task does not resemble that of its best-fit Bayesian model, which has a u-shaped beta distribution Beta(0.775,0.775). In general, the Bayesian model is a good fit to participant behavior in the six-item task, but does not account very well for participant behavior in the one-item task.

[6]Both of the empirical transition matrices are ergodic.

A close examination of the model's transition matrices and stationary distributions shows that probability matching behavior with a low standard deviation is not within this model's range of behavior.

Discussion

We have shown that learning a single versus multiple frequencies modulates participants' regularization behavior in a non-linguistic task. When participants tracked the frequency associated with a single item, they probability matched; reproducing the variation they had observed with high fidelity. However, when tracking multiple frequencies concurrently, participants regularized their responses, usually by overproducing the most common variant.

A beta-binomial Bayesian sampler model was fit to the results of each task and showed a stronger prior bias toward regularization in the six-item task than in the one-item task. Strictly speaking, the prior represents the inductive bias of the learner, and participants should come to a marble-drawing task with a particular expectation about the ratios of marbles in containers, regardless of the difficultly of the task. The fact that we find different best-fit priors according to different task demands means that we are not revealing the inductive bias of our participants, per se, but a composite picture that characterizes more than one cognitive constraint. At least one constraint that is sensitive to task demands should be added to the model, such as a memory constraint that disproportionally forgets lower-frequency observations. Such an addition could free up the prior to more accurately reflect participants' inductive bias. This raises a point of caution in comparing inductive biases across domains without controlling for task demands, since task demands can modulate bias strengths.

Our modeling results also suggest that human probability matching and regularization behavior do not lie on a simple continuum that can be captured by the prior alone. Although the Bayesian model accounted well for our participants' regularization behavior, it failed to account for the restricted variance of probability matching. Participants may be trying to produce a representative sample of draws, where the most likely response is the training ratio itself. Such a parameter might lead to high-fidelity reproduction of the training proportion under low memory constraints only.

If memory constraints are the cause of the regularization bias revealed when learning the frequencies of marbles in several containers, then this same domain-general factor may be the cause of regularization in tasks naturally characterized by concurrent frequency learning, such as language learning.

Acknowledgements

Special thanks to Tom Griffiths and Luke Maurits for feedback. This research was supported by the University of Edinburgh's College Studentship, the SORSAS award, and the Engineering and Physical Sciences Research Council.

References

Becker, A., & Veenstra, T. (2003). The survival of inflectional morphology in French-related creoles. *Studies in Second Language Acquisition, 25*, 283-306.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences, 7*, 173-221.

Chambers, J., Trudgill, P., & Schilling-Estes, N. (2003). *The handbook of language variation and change*. Blackwell, Malden, MA.

DeGraaff, M. (1999). Creolization, language change, and language acquisition: an epilogue. In M. DeGraaf (Ed.) *Language creation and language change: creolization, diachrony, and development.* Cambridge, MA: MIT Press.

Gardner, A. (1957). Probability-learning with two and three choices. *The American Journal of Psychology, 70(2)*, 174-185.

Hudson, C., & Newport, E. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development, 1(2)*, 151195.

Hudson Kam, C., & Newport, E. (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive Psychology, 59*, 30-66.

Hudson Kam, C., & Chang, A. (2009). Investigating the cause of language regularization in adults: memory constraints or learning effects? *Journal of Experimental Psychology, 35(3)*, 815-821.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science, 31*, 441480.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: sample size and the perception of correlation. *Journal of Experimental Psychology, 126(3)*, 278-287.

Lumsden, J. S. (1999). Language acquisition and creolization. In M. Degraaf (Ed.) *Language creation and language change: creolization, diachrony, and development.* Cambridge, MA: MIT Press.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11-28.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Preprint submitted to Elsevier*.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition, 111*, 317-328.

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proc. R. Soc. B, 277*, 429-436.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274(5294)*, 1926-1928.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition 116*, 444-449.

Bibliography

Alexander, R. D. (1980). *Darwinism and human affairs*. Pitman London.

Baddeley, A. (1966). The capacity for generating information by randomization. *The Quarterly journal of experimental psychology*, 18(2):119–29.

Bar-Hillel, M. and Wagenaar, W. a. (1991). The perception of randomness. *Advances in Applied Mathematics*, 12(4):428–454.

Bardsley, N. and Mehta, J. (2010). Explaining focal points: Cognitive hierarchy theory versus team reasoning. *The Economic Journal*, 120:40–79.

Barrett, J. L. and Nyhof, M. A. (2001). Spreading non-natural concepts: The role of intuitive conceptual structures in memory and transmission of cultural materials. *Journal of Cognition and Culture*, 1(1):69–100.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4*.

Bayes, M. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (1683-1775)*, pages 370–418.

Becker, A. and Veenstra, T. (2003). The survival of inflectional morphology in French-related creoles. *Studies in Second Language Acquisition*, 25(02):283–306.

Bentley, R. A. (2008). Random drift versus selection in academic vocabulary: An evolutionary analysis of published keywords. *PloS one*, 3(8):e3057.

Bentley, R. A., Hahn, M. W., and Shennan, S. J. (2004). Random drift and culture change. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1547):1443–50.

Bentley, R. A. and Shennan, S. J. (2003). Cultural transmission and stochastic network growth. *American Antiquity*, pages 459–485.

Berko, J. (1958). *The child's learning of English morphology*. PhD thesis, Radcliffe College.

Bickerton, D. (1981). *Roots of Language*. Karoma Publishers, Ann Arbor, MI.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and brain sciences*, 7(02):173–188.

Binford, L. R. (1968). *New perspectives in archeology*. Aldine Publishing Co.

Boas, M. L. (1983). *Mathematical methods in the physical sciences*. John Wiley & Sons., Inc.

Bookstein, A. (1990). Informetric distributions, Part I: Unified overview. *JASIS*, 41(5):368–375.

Boyd, R. and Richerson, P. (1985). *Culture and the evolutionary process*. The University of Chicago Press.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, 8(1):25–54.

Brugger, P. (1997). Variables that influence the generation of random sequences: an update. *Perceptual and Motor Skills*, pages 627–661.

Burkett, D. and Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. In Smith, A. D. M., Schouwtra, M., de Boer, B., and Smith, K., editors, *The Evolution of Language: Proceedings of the 8th International Conference*, pages 58–65. Singapore: World Scientific.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Busemeyer, J. R., Byun, E., Delosh, E. L., and McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In Lamberts, K. and Shanks, D. R., editors, *Knowledge, concepts, and categories: Studies in cognition*, pages 408–437. The MIT Press.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*, volume 9. John Benjamins Publishing.

Caldwell, C. A. and Smith, K. (2012). Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PloS one*, 7(8):e43807.

Campbell, D. T. (1974). Evolutionary epistemology. In Schilpp, P. A., editor, *In The philosophy of Karl Popper*, pages 413–463. La Salle, IL: Open Court.

Cavalli-Sforza, L. and Feldman, M. W. (1973). Models for cultural inheritance i. group mean and within group variation. *Theoretical population biology*, 4(1):42–55.

Cavalli-Sforza, L. L. and Feldman, M. W. (1981). Cultural transmission and evolution: a quantitative approach. *Monographs in population biology*, 16:1–388.

Chambers, J. K. and Schilling-Estes, N. (2013). *The handbook of language variation and change*, volume 80. John Wiley & Sons.

Chater, N., Reali, F., and Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences*, 106(4):1015–1020.

Chin, S. L. and Kersten, A. W. (2009). *The application of the less is more hypothesis in foreign language learning*. PhD thesis, Florida Atlantic University.

Chomsky, N. (1965). Aspects of the theory of syntax Cambridge. *Multilingual Matters: MIT Press*.

Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*, volume 6. MIT press.

Clarke, B. C. (1979). The evolution of genetic diversity. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 205(1161):453–474.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Cloutier, S., Newberry, R. C., Honda, K., and Alldredge, J. R. (2002). Cannibalistic behaviour spread by social learning. *Animal Behaviour*, 63(6):1153–1162.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Cornish, H., Tamariz, M., and Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: What experiments can tell us. *Language Learning*, 59(s1):187–205.

Cotton, J. W. and Rechtschaffen, A. (1958). Replication report: Two-and three-choice verbal-conditioning phenomena. *Journal of Experimental Psychology*, 56(1):96.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley.

Croft, W. (2003). *Typology and universals*. Cambridge University Press.

Cronk, L. (1999). *That complex whole: Culture and the evolution of human behavior*. Westview Press.

Crow, J. F. and Kimura, M. (1970). *An introduction to population genetics theory*. New York, Evanston and London: Harper & Row, Publishers.

Culbertson, J., Smolensky, P., and Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122(3):306–29.

Curio, E., Ernst, U., and Vieth, W. (1978). Cultural transmission of enemy recognition: one function of mobbing. *Science*, 202(4370):899–901.

Cutler, A., Hawkins, J. A., and Gilligan, G. (1985). The suffixing preference: a processing explanation. *Linguistics*, 23(5):723–758.

Darwin, C. (1874). *The Descent of Man*.

Davies, N. B., Krebs, J. R., and West, S. A. (2012). *An introduction to behavioural ecology*. John Wiley & Sons.

Dawkins, R. (1976). *The selfish gene*. Oxford University Press.

Dawkins, R. (1999). The extended phenotype: the long reach of the gene. *Interdisciplinary Science Reviews*.

Dediu, D. (2009). Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise? *Journal of theoretical biology*, 259(3):552–561.

Deese, J. and Kaufman, R. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of experimental psychology*, 54(3):180–187.

DeGraff, M. (1999). Creolization, language change and language acquisition: An epilogue. In DeGraff, M., editor, *Language creation and language change: Creolization, diachrony, and development*, pages 473–543. MIT Press, Cambridge, MA.

Derks, P. L. and Paclisanu, M. I. (1967). Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology*, 73(2):278.

Detambel, M. H. (1955). A test of a model for multiple-choice behavior. *Journal of Experimental Psychology*, 49(2):97.

Dobzhansky, T. G. (1951). *Genetics and the Origin of Species*, volume 11. Columbia University Press.

Dougherty, M. R. P. and Hunter, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31(6):968–982.

Du Bois, J. W. (1987). The discourse basis of ergativity. *Language*, pages 805–855.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology*, 62(4):385.

Efferson, C., Lalive, R., Richerson, P. J., McElreath, R., and Lubell, M. (2008). Conformists and mavericks: the empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior*, 29(1):56–64.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26.

Elman, J. L. (1998). *Rethinking innateness: A connectionist perspective on development*, volume 10. MIT press.

Estoup, J. B. (1916). Gammes sténographiques: méthode et exercices pour l'acquisition de la vitesse.

Ewens, W. (2004). *Mathematical Population Genetics 1: I. Theoretical Introduction*, volume 27. Springer Science & Business Media.

Fay, N. and Ellison, T. M. (2013). The cultural evolution of human communication systems in different sized populations: usability trumps learnability. *PloS one*, 8(8):e71781.

Fay, N., Garrod, S., and Roberts, L. (2008). The fitness and functionality of culturally evolved communication systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3553–3561.

Fay, N., Garrod, S., Roberts, L., and Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, 34(3):351–386.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407(6804):630–633.

Feldman, M. W. and Laland, K. N. (1996). Gene-culture coevolutionary theory. *Trends in Ecology & Evolution*, 11(11):453–457.

Felsenstein, J. (2005). *Theoretical Evolutionary Genetics*. University of Washington, Seattle.

Ferdinand, V., Thompson, B., Kirby, S., and Smith, K. (2013). Regularization behavior in a non-linguistic domain. *Knauff, M., Sebanz, N., Pauen, M., and Wachsmuth, I.(Eds.) Bielefeld University Proceedings of the 35th Annual Cognitive Science Society*.

Ferdinand, V. and Zuidema, W. (2009). Thomas' theorem meets Bayes' rule: a model of the iterated learning of language. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 1786–1791.

Ferrer i Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791.

Fiser, J. and Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, 12(6):499–504.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford: Clarendon Press.

Flynn, E. (2008). Investigating children as cultural magnets: do young children transmit redundant information along diffusion chains? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3541–3551.

Flynn, E. and Whiten, A. (2008). Cultural transmission of tool use in young children: A diffusion chain study. *Social Development*, 17(3):699–718.

Futuyma, D. J. (1998). *Evolutionary biology*. Sinauer.

Gaissmaier, W. and Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, 109(3):416–22.

Gajdon, G. K., Fijn, N., and Huber, L. (2004). Testing social learning in a wild mountain parrot, the kea (Nestor notabilis). *Animal Learning & Behavior*, 32(1):62–71.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive science*, 29(5):737–767.

Galantucci, B., Kroos, C., and Rhodes, T. (2010). The effects of rapidity of fading on communication systems. *Interaction Studies*, 11(1):100–111.

Gardner, R. (1958). Multi-choice decision behavior. *American Journal of Psychology*, 71.

Gardner, R. A. (1957). Probability-Learning with two and three choices. *The American Journal of Psychology*, 70(2):174–185.

Garrod, S., Fay, N., Lee, J., Oberlander, J., and MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31(6):961–987.

Garrod, S., Fay, N., Rogers, S., Walker, B., and Swoboda, N. (2010). Can iterated learning explain the emergence of graphical symbols? *Interaction Studies*, 11(1):33–50.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, R. (1998). Domain specificity in cognitive development: universals and nonuniversals. *Advances in psychological science*.

Goldowsky, B. N. and Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: the less is more hypothesis. In *The proceedings of the 24th annual child language research forum*, pages 124–138.

Goldstone, R. L. and Gureckis, T. M. (2009). Collective behavior. *Topics in Cognitive Science*, 1(3):412–438.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *The Journal of Machine Learning Research*, 12:2335–2382.

Goldwater, S., Johnson, M., and Griffiths, T. L. (2005). Interpolating between types and tokens by estimating power-law generators. In *Advances in neural information processing systems*, pages 459–466.

Goodman, N. (1954). Fact, Fiction, & Forecast. *Cambridge, Mass.: Harvard University Press*.

Grassberger, P. (2003). Entropy estimates from insufficient samplings. *arXiv preprint physics/0307138*.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *In J. Greenberg, ed., Universals of Language. 73-113. Cambridge, MA.*

Grether, D. M. (1992). Testing bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, 17(1):31–57.

Griffiths, T. and Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31:441–480.

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

Grimshaw, J. (1997). Projection, heads, and optimality. *Linguistic inquiry*, pages 373–422.

Gureckis, T. M. and Goldstone, R. L. (2009). How you named your child: Understanding the relationship between individual decision making and collective outcomes. *Topics in Cognitive Science*, 1(4):651–674.

Hahn, M. W. and Bentley, R. A. (2003). Drift as a mechanism for cultural change: an example from baby names. *Proceedings. Biological sciences / The Royal Society*, 270 Suppl:S120–S123.

Hahn, U. and Warren, P. a. (2009). Perceptions of randomness: why three heads are better than four. *Psychological review*, 116(2):454–61.

Hasher, L. and Zacks, R. (1979). Automatic and effortful processes in memory. *Journal of experimental psychology: General*.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the experimental analysis of behavior*, 13(2):243–266.

Herzog, H. A., Bentley, R. A., and Hahn, M. W. (2004). Random drift and large shifts in popularity of dog breeds. *Proceedings. Biological sciences / The Royal Society*, 271 Suppl:S353–S356.

Hinton, G. E. and Nowlan, S. J. (1987). How learning can guide evolution. *Complex systems*, 1(3):495–502.

Horner, V., Whiten, A., Flynn, E., and de Waal, F. B. M. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences*, 103(37):13878–13883.

Hudson Kam, C. L. and Chang, A. (2009). Investigating the cause of language regularization in adults: memory constraints or learning effects? *Journal of experimental psychology. Learning, memory, and cognition*, 35(3):815–21.

Hudson Kam, C. L. and Newport, E. L. (2005). Regularizing unpredictable variation : The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195.

Hudson Kam, C. L. and Newport, E. L. (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive psychology*, 59(1):30–66.

Hume, D. (1793). *A Treatise of Human Nature*.

Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77(2):187–222.

Hurford, J. R. (2002). Expression/induction models of language evolution: Dimensions and issues. In Briscoe, E. J., editor, *Linguistic evolution through language acquisition: Formal and computational models.* Cambridge University Press, Cambridge UK.

Hurford, J. R. (2003). The language mosaic and its evolution. *Studies in the Evolution of Language*, 3:38–57.

Hurford, J. R. (2011). *The origins of grammar: Language in the light of evolution II*, volume 2. Oxford University Press.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge University Press.

Jescheniak, J. D. and Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4):824.

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., and Tyler-Smith, C. (2013). *Human evolutionary genetics: origins, peoples & disease*. Garland Science.

Kaas, R. and Buhrman, J. (1980). Mean, median and mode in binomial distributions. *Statistica Neerlandica*, pages 13–18.

Kahneman, D. and Tversky, A. (1972). Subjective probability : A judgment of representativeness. *Cognitive psychology*, 3:430–454.

Kalish, M. L., Griffiths, T. L., and Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2):288–294.

Kameda, T. and Nakanishi, D. (2002). Cost–benefit analysis of social/cultural learning in a nonstationary uncertain environment: An evolutionary simulation and an experiment with human subjects. *Evolution and Human Behavior*, 23(5):373–393.

Kareev, Y., Lieberman, I., and Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology*, 126(3):278–287.

Keesing, R. M. (1974). Theories of Culture. *Annual Review of Anthropology*, 3:73–97.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, 10(3):307–321.

Keynes, J. (1921). *A treatise on probability*. Macmillan and Co. Limited.

Kirby, S. (1999). *Function, selection and innateness: The emergence of language universals*. PhD thesis, University of Edinburgh.

Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In Knight, C., Studdert-Kennedy, M., and Hurford, R., editors, *The evolutionary emergence of language*. Cambridge University Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *Evolutionary Computation, IEEE Transactions on*, 5(2):102–110.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, E., editor, *Linguistic evolution through language acquisition: Formal and computational models.*, pages 173–203. Cambridge University Press.

Kirby, S. (2013). Transitions: The evolution of linguistic replicators. In *The Language Phenomenon*, pages 121–138. Springer.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, 105(31):10681–10686.

Kirby, S., Dowman, M., and Griffiths, T. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245.

Kirby, S., Griffiths, T., and Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114.

Kirk, K. L. and Bitterman, M. E. (1965). Probability-learning by the turtle. *Science*, 148(3676):1484–1485.

Komarova, N. L., Niyogi, P., and Nowak, M. A. (2001). The evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology*, 209(1):43–59.

Kroeber, A. and Kluckhohn, C. (1952). Culture: A critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University.*

Lachmann-Tarkhanov, M. and Sarkar, S. (1994). The alternative fitness sets which preserve allele trajectories: a general treatment. *Genetics*, 138(4):1323–1330.

Laland, K. N. (2004). Social learning strategies. *Animal Learning & Behavior*, 32(1):4–14.

Laland, K. N. and Plotkin, H. C. (1993). Social transmission of food preferences among Norway rats by marking of food sites and by gustatory contact. *Animal Learning & Behavior*, 21(1):35–41.

Laland, K. N. and Williams, K. (1997). Shoaling generates social learning of foraging information in guppies. *Animal Behaviour*, 53(6):1161–1169.

Langen, T. A. (1996). Social learning of a novel foraging skill by white-throated magpie-jays (*Calocitta formosa, Corvidae*): A field experiment. *Ethology*, 102(1):157–166.

Lewandowsky, S., Griffiths, T. L., and Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, 33(6):969–998.

Lewis, D. (1969). *Convention: A philosophical study.* John Wiley & Sons.

Lewontin, R. C. (1970). The units of selection. *Annual Review of Ecology and Systematics*, pages 1–18.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., and Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716.

Lloyd, G. E. R. (1968). *Aristotle: the growth and structure of his thought.* Cambridge University Press Cambridge.

Ludden, D. and Gupta, P. (2000). Zen in the art of language acquisition: Statistical learning and the less is more hypothesis. In *22nd Annual Conference of the Cognitive Science Society.* Citeseer.

Lumsden, C. J. and Wilson, E. O. (1981). *The coevolutionary process.* World Scientific.

Lumsden, J. S. (1999). Language acquisition and creolization. In DeGraaf, M., editor, *Language creation and language change: Creolization, diachrony, and development*, chapter Language a. Cambridge, MA: MIT Press.

Marantz, A. (1995). The minimalist program. *Government and binding theory and the minimalist program*, pages 351–382.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., and Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4):i–178.

Mars, R. B., Shea, N. J., Kolling, N., and Rushworth, M. F. S. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly journal of experimental psychology (2006)*, 65(2):252–67.

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101—-B111.

Maynard Smith, J. (1989). *Evolutionary genetics.* Oxford University Press.

Mayr, E. (1982). *The growth of biological thought: diversity, evolution, and inheritance.* Harvard University Press.

McCormack, P. (1959). Spatial generalization and probability-learning in a five-choice situation. *The American Journal of Psychology*, pages 135–138.

McElreath, R., Bell, A. V., Efferson, C., Lubell, M., Richerson, P. J., and Waring, T. (2008). Beyond existence and aiming outside the laboratory: Estimating frequency-dependent and pay-off-biased social learning strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3515–3528.

McElreath, R., Lubell, M., Richerson, P. J., Waring, T. M., Baum, W., Edsten, E., Efferson, C., and Paciotti, B. (2005). Applying evolutionary models to the laboratory study of social learning. *Evolution and Human Behavior*, 26(6):483–508.

McEwen, F., Happe, F., Bolton, P., Rijsdijk, F., Ronald, A., Dworzynski, K., and Plomin, R. (2007). Origins of individual differences in imitation: Links with language, pretend play, and socially insightful behavior in two-year-old twins. *Child development*, 78(2):474–492.

Mehta, J., Starmer, C., and Sugden, R. (1994a). Focal points in pure coordination games: an experimental investigation. *Theory and Decision*, pages 163–185.

Mehta, J., Starmer, C., and Sugden, R. (1994b). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, 84(3):658–673.

Mendel, G. (1866). Experiments on plant hybrids (English translation). In C Stern, E. S., editor, *The Origin of Genetics: A Mendel Source Book*, pages 1–48. San Francisco: WH Freeman.

Menzel, E. W. (1973). Further observations on the use of ladders in a group of young chimpanzees. *Folia Primatologica*, 19(6):450–457.

Mesoudi, A. (2011). *Cultural Evolution.* The University of Chicago Press.

Mesoudi, A. and Lycett, S. J. (2009). Random copying, frequency-dependent copying and culture change. *Evolution and Human Behavior*, 30(1):41–48.

Mesoudi, A. and O'Brien, M. J. (2008). The cultural transmission of Great Basin projectile-point technology I: an experimental simulation. *American Antiquity*, pages 3–28.

Mesoudi, A. and Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of cognition and culture*, 4(1):1–24.

Mesoudi, A., Whiten, A., and Dunbar, R. (2006a). A bias for social information in human cultural transmission. *British Journal of Psychology*, 97(3):405–423.

Mesoudi, A., Whiten, A., and Laland, K. N. (2004). Perspective: is human cultural evolution Darwinian? Evidence reviewed from the perspective of The Origin of Species. *Evolution*, 58(1):1–11.

Mesoudi, A., Whiten, A., and Laland, K. N. (2006b). Towards a unified science of cultural evolution. *The Behavioral and brain sciences*, 29(4):329–47; discussion 347–83.

Meyerhoff, M. (2000). The emergence of creole subject–verb agreement and the licensing of null subjects. *Language Variation and Change*, 12:203–230.

Miller, M. B. and Valsangkar-Smyth, M. (2005). Probability matching in the right hemisphere. *Brain and cognition*, 57(2):165–7.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251.

Mokkonen, M., Kokko, H., Koskela, E., Lehtonen, J., Mappes, T., Martiskainen, H., and Mills, S. C. (2011). Negative frequency-dependent selection of sexually antagonistic alleles in Myodes glareolus. *Science*, 334(6058):972–974.

Moran, P. a. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(01):60–71.

Müller, M. (1870). Darwinism tested by the science of language (translated from the German of Professor August Schleicher). *Nature*, 1(10):256–259.

Murphy, K. (2001). Learning Bayes net structure from sparse data sets. *Technical report, Comp. Sci. Div., UC Berkeley*.

Myers, J. L. (1976). Probability learning and sequence learning. *Handbook of learning and cognitive processes: Approaches to human learning and motivation*, 3:171–205.

Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49.

Neal, D. (2004). *Introduction to population biology*. Cambridge University Press.

Neiman, F. D. (1995). Stylistic variation in evolutionary perspective: Inferences from decorative diversity and interassemblage distance in Illinois woodland ceramic assemblages. *American Antiquity*, 60:7–36.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive science*, 14(1):11–28.

Newport, E. L. and Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2):127–162.

Norris, J. R. (2008). Markov chains.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., and Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & cognition*, 22(3):352–369.

Nowak, M. A. and Komarova, N. L. (2001). Towards an evolutionary theory of language. *Trends in cognitive sciences*, 5(7):288–295.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2001). Evolution of universal grammar. *Science*, 291(5501):114–118.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617.

O'Brien, M. J., Darwent, J., and Lyman, R. L. (2001). Cladistics is useful for reconstructing archaeological phylogenies: Palaeoindian points from the southeastern United States. *Journal of Archaeological Science*, 28(10):1115–1136.

Odling-Smee, F., Laland, K., and Feldman, M. (2003). *Niche construction: the neglected process in evolution*. Princeton University Press.

Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive behavior*, 7(3-4):371–383.

Osborne, M. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

Papoulis, A. (1984). Brownian movement and markoff processes. *Ch*, 15:515–553.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, pages 1–60.

Perfors, A. (in preparation). Adult regularization of inconsistent input depends on pragmatic factors.

Plotkin, H. C. (1994). *Darwin machines and the nature of knowledge*. London, UK: Penguin.

Popper, K. R. (1959). The logic of scientific discovery. *London: Hutchinson*.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rafferty, A. N. and Griffiths, T. L. (2010). Optimal language learning : The importance of starting representative. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Raiffa, H. and Schlaifer (1961). *Applied statistical decision theory*. Division of Research, Harvard Business School, Cambridge, MA.

Reali, F. and Griffiths, T. L. (2009). The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. *Cognition*, 111(3):317–28.

Reali, F. and Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings. Biological sciences / The Royal Society*, 277(1680):429–36.

Redington, M. and Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of experimental psychology: general*, 125(2):123.

Richerson, P. J. and Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Rick, S. and Weber, R. a. (2010). Meaningful learning and transfer of learning in games played repeatedly without feedback. *Games and Economic Behavior*, 68(2):716–730.

Rieseberg, L. H. (1997). Hybrid origins of plant species. *Annual Review of Ecology and Systematics*, pages 359–389.

Rigden, C. (1999). 'The eye of the beholder'-designing for colour-blind users. *British Telecommunications Engineering*.

Rosenthal, J. S. (1995). Convergence rates for Markov chains. *Siam Review*, 37(3):387–405.

Rumelhart, D. E. and McClelland, J. L. (1985). On learning the past tenses of English verbs. *ICS Report 8507*.

Saffran, J. R. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science*, 12(4):110–114.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Sankoff, G. (1979). The Genesis of a Language. In Hill, K. C., editor, *The genesis of language*, pages 23–47. Karoma Publishers, Ann Arbor, MI.

Sapir, E. (1921). *Language: An introduction to the study of speech*. New York, Harcourt, Brace and World.

Saussure, F. d. (1966). *Course in general linguistics*. McGraw Hill, New York.

Schelling, T. (1960). The Strategy of Conflict. *Cambridge, Mass*.

Schilling-Estes, N. and Wolfram, W. (1994). Convergent explanation and alternative regularization patterns: Were/weren't leveling in a vernacular English variety. *Language variation and change*, 6(03):273–302.

Schotter, A. and Sopher, B. (2003). Social learning and coordination conventions in intergenerational games: An experimental study. *Journal of Political Economy*, 111(3):498–529.

Scott-Phillips, T. C., Dickins, T. E., and West, S. A. (2011). Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, 6(1):38–47.

Scott-Phillips, T. C. and Kirby, S. (2010). Language evolution in the laboratory. *Trends in cognitive sciences*, 14(9):411–417.

Scott-Phillips, T. C., Kirby, S., and Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113(2):226–33.

Senghas, A. (2000). The development of early spatial morphology in Nicaraguan Sign Language. In *Proceedings of the 24th Annual Boston University Conference on Language Development*, pages 696–707.

Senghas, A. and Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological Science*, 12(4):323–328.

Senghas, A., Coppola, M., Newport, E. L., and Supalla, T. (1997). Argument structure in Nicaraguan Sign Language: The emergence of grammatical devices. In *Proceedings of the 21st Annual Boston University Conference on Language Development*, pages 550–561. Cascadilla Press Boston, MA.

Shanks, D. R., Tunney, R. J., and McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, 15(3):233–250.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, 75(13):1–42.

Silvey, C., Kirby, S., and Smith, K. (2014). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive science*.

Singh, K. and Xie, M. (2008). Bootstrap: a statistical method. *Unpublished Working Paper. Rutgers University,¡ http://www. stat. rutgers. edu/home/mxie/RCPapers/bootstrap. pdf*.

Singleton, J. L. and Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, 49(4):370–407.

Smith, J., Durham, M., and Fortune, L. (2007). "Mam, my trousers is fa'in doon!": Community, caregiver, and child in the acquisition of variation in a Scottish dialect. *Language Variation and Change*, 19:63–99.

Smith, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Science*, 14(1):65–84.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 697–702. Citeseer.

Smith, K. and Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1509):3591–603.

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

Smith, K. and Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3):444–9.

Sober, E. (1984). *The nature of selection: Evolutionary theory in philosophical focus*. University of Chicago Press.

Steels, L. (1999). The talking heads experiment. *Laboratorium*.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in cognitive sciences*, 7(7):308–312.

Stewart, W. J. (1994). *Introduction to the numerical solution of Markov chains*, volume 41. Princeton University Press Princeton.

Sugden, R. (1995). A theory of focal points. *The Economic Journal*, pages 533–550.

Sumita, K., Kitahara-Frisch, J., and Norikoshi, K. (1985). The acquisition of stone-tool use in captive chimpanzees. *Primates*, 26(2):168–181.

Symons, D. (1979). *The evolution of human sexuality*. Oxford University Press.

Szathmary, E. and Maynard Smith, J. (2004). *The major transitions in evolution*. Oxford University Press.

Tamariz, M. and Smith, A. D. (2008). Regularity in mappings between signals and meanings. In *The Evolution of Language: Proceedings of the 7th international conference*, pages 315–322.

Tenenbaum, J. B. and Griffiths, T. L. (2001). The rational basis of representativeness. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Thompson, B., Smith, K., and Kirby, S. (2012). Cultural evolution renders linguistic nativism implausible. In T. Scott-Phillips, M. T. . E. C., editor, *The Evolution of Language: Proceedings of the 9th International Conference*.

Tullo, C. and Hurford, J. (2003). Modelling Zipfian distributions in language. In *Proceedings of language evolution and computation workshop/course at ESSLLI*, pages 62–75.

Tune, G. (1964). A brief survey of variables that influence random-generation. *Perceptual and motor skills*.

Unturbe, J. and Corominas, J. (2007). Probability matching involves rule-generating ability: A neuropsychological mechanism dealing with probabilities. *Neuropsychology*, 21(5):621–30.

van Trijp, R. (2013). Linguistic assessment criteria for explaining language change: A case study on syncretism in German definite articles. *Language Dynamics and Change*, 3(1):105–132.

Vanderschraaf, P. (2014). *Learning and coordination: Inductive deliberation, equilibrium and convention*. Routledge.

Vennemann, T. (1975). An explanation of drift. *Word Order and Word Order Change*.

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, 4.

Vickers, J. (2014). The problem of induction. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014 edition.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial intelligence*, 167(1):206–242.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107(2):729–42.

Vulkan, N. (2000). An economists perspective on probability matching. *Journal of economic surveys*, 14(1):101–118.

Weber, R. (2003). 'Learning' with no feedback in a competitive guessing game. *Games and Economic Behavior*.

Wegener (1912). *The Origin of Continents*. Geol. Rundsch. 3.

Weir, M. W. (1964). Developmental changes in problem-solving strategies. *Psychological Review*, 71(6):473.

Weir, M. W. (1972). Probability performance: Reinforcement procedure and number of alternatives. *The American Journal of Psychology*, pages 261–270.

Whiten, A. and Mesoudi, A. (2008). Establishing an experimental science of culture: animal social diffusion experiments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3477–3488.

Williams, D. C. (1963). *The ground of induction*. Cambridge, MA, Harvard University Press.

Wilson, D. S. (1975). A theory of group selection. *Proceedings of the national academy of sciences*, 72(1):143–146.

Wilson, R. and Rhodes, C. (1997). Leadership and credibility in n-person coordination games. *Journal of Conflict Resolution*, 41(6):767–791.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf]*, page 42.

Wittig, M. A. and Weir, M. W. (1971). The role of reinforcement procedure in children's probability learning as a function of age and number of response alternatives. *Journal of Experimental Child Psychology*, 12(2):228–239.

Wolford, G., Miller, M. B., and Gazzaniga, M. (2000). The left hemisphere's role in hypothesis formation. *The Journal of Neuroscience*.

Wolford, G., Newman, S. E., Miller, M. B., and Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 58(4):221–228.

Wolpert, D. H. and DeDeo, S. (2013). Estimating functions of distributions defined over spaces of unknown size. *Entropy*, 15(11):4668–4699.

Wonnacott, E. and Newport, E. (2005). Novelty and regularization: The effect of novel instances on rule formation. In Brugos, A., Clark-Cotton, M. R., and Ha, S., editors, {*BUCLD 29*}*: Proceedings of the 29th Annual Boston University Conference on Language Development*, Boston, MA. Cascadilla Press.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):97–159.

Wright, S. (1969). *Evolution and the genetics of populations: The theory of gene frequencies*, volume 2. University of Chicago Press.

Xu, J., Dowman, M., and Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, 280(1758):20123073.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language.* Harvard University Press.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology.* Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press.

Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. *Advances in neural information processing systems*, pages 51–58.