12-2014

Unnatural pedagogy : a computational analysis of children's learning to learn from other people.

Baxter S. Eaves 1984-
University of Louisville

Unnatural Pedagogy: A computational analysis of children's learning to learn from other people

By

Baxter S. Eaves Jr.
B.A., University of Louisville, 2011
M.S., University of Louisville, 2013

A Dissertation Submitted to the Faculty of the College of Arts and Sciences of the University of Louisville in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Department of Psychological and Brain Sciences
University of Louisville
Louisville, Ky

December 2014

Unnatural Pedagogy: A computational analysis of children's learning to learn from other people

By

Baxter S. Eaves Jr.
B.A., University of Louisville, 2011
M.S., University of Louisville, 2013

A Dissertation approved on

20 November 2014

by the following Dissertation Comittee:

Patrick Shafto, Ph.D.

Judith Danovitch, Ph.D.

Olfa Nasraoui, Ph.D.

Nicholaus Noles, Ph.D.

John Pani, Ph.D.

Dedication

To Baxter Eaves Sr.

One could not ask for a more loving, supportive, well–mustachioed dad.

Acknowledgments

First and foremost, I would like to express my sincerest appreciation to Patrick Shafto, whose support and guidance over the years has made all of this work possible. I am indebted to Dr. Shafto not only for the effort he has put into developing this work, but for the effort he has put into developing me as a scientist and as a professional.

Dan Navarro and Amy Perfors were major collaborators on the work that forms the foundation of this work and Thomas Griffiths and Naomi Feldman were instrumental to the work on infant–directed speech. I would also like to thank my lab comrades: Nick Searcy, Kelley Durkin, Asheley Landrum, and Campbell Rightmyer who have provided valuable feedback on the many projects I have conducted over the years that comprise this work. I am grateful to Vikash Mansinghka who has helped me to improve my mathematical understanding and who has graciously provided me with computational resources that have allowed me to conduct this work at an unexpected pace. I would like to thank my committee members Judith Danovitch, Olfa Nasraoui, Nicholaus Noles, and John Pani who have provided valuable feedback and who have helped me to consider my research from other perspectives.

Finally, I would like to thank Alaina Mitchell for her companionship, compassion, support, and stoic endurance of my nonsense during the writing of this document.

Abstract

Unnatural Pedagogy: A computational analysis of children's learning to learn from other people

Baxter S. Eaves Jr.

20 November 2014

Infants rely on others for much of what they learn. People are a ready source of quick information, but people produce data differently than the world. Data from a person are a result of that person's knowledgeability and intentions. People may produce inaccurate or misleading data. On the other hand, if a person is knowledgeable about the world and intends to teach, that person may produce data that are more useful than simply accurate data: data that are pedagogical. This idea that people have special innate methods for efficient information transfer lies at the heart of recent proposals regarding what makes humans such powerful knowledge accumulators. These innate assumptions result in developmental patterns observed in epistemic trust research. This research seeks to create a computational account of the development of these abilities. We argue that pedagogy is not innate, but rather that people learn to learn from others. We employ novel computational models to show that there is sufficient data early on from which infants may learn that people choose data pedagogically, that the development of children's epistemic trust is primarily a result of their decreasing beliefs that all informants are helpful, and that innate pedagogy would not lead to more rapid learning. We connect results from the pedagogy and epistemic trust literatures across tasks and development, showing that these are different manifestations of the same underlying abilities, and show that pedagogy need not be innate to have powerful implications for learning.

TABLE OF CONTENTS

Acknowledgments **iv**

Abstract **v**

List of Tables **ix**

List of Figures **x**

1 Introduction **1**

1.1 Pedagogical data selection 2

1.2 Natural Pedagogy 3

1.3 Statistical learning 6

1.4 Learning to learn socially 7

2 A computational argument for the early availability of pedagogically optimized data* **10**

2.1 Introduction 10

2.2 Model 11

2.3 Methods 13

2.4 Results 15

2.5 Discussion 18

2.5.1 Limitations 20

2.5.2 Conclusion 20

2.6 Detailed model specification 20

2.6.1 Learner model 20

2.6.2 Teacher model 23

2.7 Summary 23

3 A probabilistic model of learning from informants 25

3.1 Probabilistic models of data selection 25

3.2 Epistemic trust model specification 27

3.3 The difficulty of deception or: *never go in against a Sicilian when death is on the line* 31

3.4 An example from word learning 33

4 Parameterizing developmental trends in the literature using computational analysis 37

4.1 Cross-categorization 38

4.2 Method 39

4.3 Analysis and results 41

5 Modeling epistemic trust across tasks and ages 47

5.1 A note on parameter fitting 47

5.2 Procedure and results 50

5.2.1 Accuracy 50

5.2.2 Relative accuracy 52

5.2.3 Familiarity 54

5.2.4 Consensus 56

5.2.5 Culture 58

5.2.6 Deceptive pointing and marking 60

5.2.7 Error magnitude 62

5.2.8 Looking time 65

5.2.9 Gaze following 66

5.3 Capturing the effect of attachment during infancy on trust 70

5.4 Discussion 75

6 An argument against the necessity of an innate pedagogical assumption 77

6.1 Methods 79

6.1.1 Alternative models 79

6.1.2 Simulating informants 80

6.1.3 Learning situations 81

6.1.4 Simulating learners 82

6.1.5 Procedure 83

6.2 Results 85

6.2.1 Learning 85

6.2.2 Learning under causal and labeling trials 88

6.3 Limitations 89

6.4 Conclusion 91

7 Conclusion 94

7.1 Future Directions 96

7.2 Conclusion 97

References 97

Appendices 106

A Monte Carlo estimation in the epistemic trust model 106

A.1 Rejection sampling 106

A.2 Gibbs sampling 107

B Expected dependence probability between cross-categorization columns 110

Curriculum Vitae 112

LIST OF TABLES

2.1 List of phonemes in International Phonetic Alphabet transcription with means and variances calculated from the data. 14

4.1 Example structure of a hypothetical prepared table used during cross-categorization 41

5.1 List of studies modeled. 48

5.2 Clusters of experiments used for grid search parameter optimization. 49

5.3 Best-fitting parameter sets for both strangers and mothers for each attachment level in Corriveau, Harris, et al. (2009). 75

6.1 Informant types used for simulations. 81

6.2 Simulation results reference. 85

LIST OF FIGURES

1.1 Random and pedagogical sampling for defining hidden rectangles. 3

2.1 Distributions of vowels along first and second formants (F1 and F2) in adult-directed speech and speech optimized for the learner. 16

2.2 Mean change in Euclidean distance between phonemes pairs from ADS to optimal data over ten independent simulations. 17

2.3 Mean change in F_1 variance, F_2 variance, and covariance ADS to optimal data over ten independent simulations. 18

3.1 Graphical representation of the model. 28

3.2 Beta distributions for a variety of parameter values. 30

4.1 Resulting dependence probability matrix from cross-categorization. 42

4.2 Conditional probability distribution of helpfulness and knowledgeability parameters given age. 45

4.3 Conditional distribution of helpfulness's balance distribution given each communication mode. 46

5.1 Model simulation results for Koenig and Harris (2005). 51

5.2 Model simulation results for Fitneva and Dunfield (2010). 53

5.3 Model simulation results for Pasquini, Corriveau, Koenig, and Harris (2007) (clustered parameter fit). 54

5.4 Model simulation results for Pasquini, Corriveau, Koenig, and Harris (2007) (best-fitting parameters). 55

5.5 Model simulation results for Corriveau and Harris (2009) 56

5.6 Model simulation results for Corriveau, Fusaro, and Harris (2009) and Chen, Corriveau, and Harris (2012) 57

5.7 Model simulation results for DiYanni, Corriveau, Nasrini, Kurkul, and Nini (under review) 59

5.8 Model simulation results for Couillard and Woodward (1999) 61

5.9 Model simulation results for Einav and E. J. Robinson (2010). 64

5.10 Model simulation results for Koenig and Echols (2003) 66

5.11 Model simulation results for Tummeltshammer, Wu, Sobel, and Kirkham (2014) 69

5.12 Distribution of unknowledgeable informant beliefs given world states for (hybrid object saliency) hybrid objects and weight function. 73

5.13 Simulated and empirical data for Corriveau, Harris, et al. (2009) 73

6.1 Models used for simulations. 80

6.2 Mean learning from 70% teachers given neutral starting beliefs and a fixed number of informants over 50 trials. 86

6.3 Mean learning from 70% teachers given neutral starting beliefs and a variable number of informants over 50 trials. 87

6.4 Mean learning from 70% teachers given strongly biased starting beliefs and a variable number of informants over 50 trials. 89

6.5 Learning given only teachers. 90

6.6 Learning rates for causal-only and word-only simulations. 93

A.1 Example Markov blanket for epistemic trust model and Gibbs sampling conditional probabilities and distributions. 109

1. INTRODUCTION

> Muad'Dib learned rapidly because his first training was in how to learn. And the first lesson of all was the basic trust that he could learn.
>
> Herbert (2003)

Newborns face an unfathomably difficult task: learn the world (James, 1890). To function in the world, infants must learn physical concepts such as gravity and dynamics as well as human concepts such as language and mathematics. Self-directed strategies are slow and cannot be used to acquire the full set of required knowledge. For example, language cannot be learned through trial-and-error and intervention is a dangerous tactic in some situations such as learning the outcome of inserting silverware into power outlets. For these things, infants rely on others. People are a ready source of quick information, but people produce data differently than the world (Dennett, 1978; Gergely, Egyed, & Király, 2007).* Data from the world adhere to the laws of nature; they are a direct result of their true generative process. Data from a person, on the other hand, are a result of that person's knowledgeability and intentions. People may be knowledgeable about the world but not willing or able to share their knowledge; they may be willing and able to share their knowledge but may posses inaccurate knowledge. In these cases, people may produce inaccurate or misleading data. On the other hand, if a person is knowledgeable about the world and intends to teach—to induce that particular hypothesis in the learner—that person may produce data that are better than accurate: data that are pedagogical (Shafto & Goodman,

*In this work we refer to *the world* in a variety of ways. The world is the set of things that do not entertain beliefs. An informant flips a light switch but the world pushes the electrons through the copper wire which causes the tungsten filament to heat and emit photons. When we refer to the *true state of the world*, we mean the correct hypothesis, e.g. that a switch turns on a light, or that a particular downy quadruped is member of the *Felis* genus and is named *Rodger*.

2008; Shafto, Goodman, & Griffiths, 2014). This idea that people have special methods for efficient information transfer lies at the heart of recent proposals regarding what makes humans such powerful knowledge accumulators (Tomasello, 2009; Csibra & Gergely, 2006; Boyd, Richerson, & Henrich, 2011; Pinker, 2010).

1.1. Pedagogical data selection

Pedagogical data result when informants are knowledgeable about the world and intend to teach learners (Csibra & Gergely, 2006; Gergely et al., 2007; Bonawitz et al., 2011). Pedagogical data are most illustrative of their hypothesis in the sense that they lead learners toward that hypothesis while leading them away from alternative hypotheses (Shafto & Goodman, 2008; Shafto et al., 2014). They are sampled with an implied meta-reasoning between learners and informants: informants expect that learners expect pedagogical data, learners expect informants to expect that they expect pedagogical data, and so on. As an illustrative example take Shafto and Goodman (2008) in which participants were to learn the size and location of invisible rectangles using positive (inside the rectangle) and negative (outside the rectangle) examples (see Figure 1.1). To fully specify a single rectangle to a naive learner, six points are required: two positive, in opposite corners, to constrain the minimum size; and four negative, one on each edge, to constrain the maximum size. Given pedagogical sampling, only two positive examples, in opposite corners, are needed. A teacher places points at opposite corners to rule out all smaller rectangles; the learner can infer that the true rectangle is the smallest rectangle to encompass both points because if the rectangle were larger the teacher would have placed the examples farther apart. In this case, a learner assuming that data are sampled pedagogically needs only one third of the amount data needed by a learner assuming that data are sampled randomly. It should also be noted that in this particular example, the number of data points needed under pedagogical sampling is invariant to the dimensionality. Only two positive examples are required to teach any n-dimensional hypercube. In this case, the data savings increase as the concept increases in complexity.

Though pedagogical data are often beneficial to learners assuming random sampling,

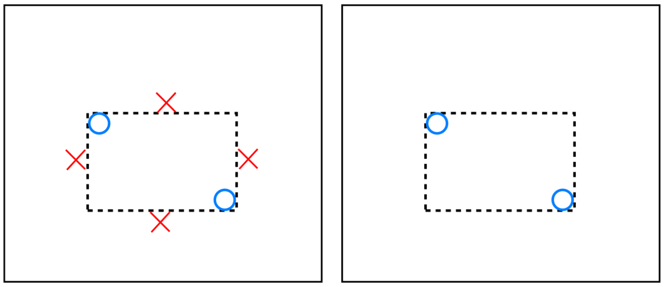


Figure 1.1: Random and pedagogical sampling for defining hidden rectangles. Blue circles represent positive examples, red X's represent negative examples, and dashed lines represent the hidden, target rectangle. *Left*) Fully specifying the target rectangle using deduction, assuming randomly-sampled data. *Right*) Fully specifying the target rectangle with pedagogically-sampled data.

to fully benefit from pedagogical data, learners must recognize when pedagogical sampling applies and must also understand the intent behind pedagogical data. For example, though pedagogically–placed negative examples will rule out all larger rectangles for a learner assuming random sampling, the learner assuming pedagogical sampling will recognize that the examples were placed to rule out smaller rectangles as well because the teacher desires to convey a single concept. Learners rely on teachers to choose specific data and teachers rely on learners to learn with this in mind.

1.2. Natural Pedagogy

A current leading theory of social learning, *Natural Pedagogy* (Csibra & Gergely, 2006) argues that humans, and only humans, are born with the innate ability to recognize, engage in, and exploit acts of pedagogy. According to Csibra and Gergely (2009) people's learning is "[. . .]supported by powerful learning mechanisms that capitalize on innate biases[. . .]" (p. 152). Natural pedagogy asserts that people have evolved the ability to infer others' goals from their actions and to learn with these goals in mind. The theory argues that people use so-called *ostensive cues* such as eye contact, pointing, and infant-directed speech (IDS, motherese)* and that people recognize that these cues signal an impending pedagogical demonstration. It is argued that ostensive cues not only capture infants' attention (a vital

*Infant-directed speech is speech to infants characterized by reduced speed, elevated pitch and affect, and unusual prosody.

part of teaching), but prime infants to receive referential information. For example, at a zoo, a mother smiles at her child, points to an animal, and says to the child, “Wow! Look at the *giraffe*.”

The idea of Natural Pedagogy has inspired a great deal of interesting research and reinterpretations of classic results. For example, Meltzoff (Meltzoff, 1988, 1995) conducted a study in which infants imitated the novel action of an adult, specifically, activating a light on a table by bending at the waist while seated and pressing it with the forehead. Infants performed the same novel, but inefficient, action to activate the light. This result was interpreted as an example of children’s imitative learning and ability to use novel strategies in novel situations. Why would children perform a difficult action, pressing with the head, when a simpler action, pressing with the hand, would produce the same outcome? The pedagogical stance suggests that because the demonstrator did not use his hands, that using the hands is not likely to produce the desired outcome because people act efficiently toward their goals. To this end Gergely, Bekkering, and Király (2002) repeated the experiment with an additional condition. In the hands-occupied condition, the demonstrator performed the head-press action while huddled shivering in a blanket. Infants who observed the hands-occupied demonstration were more likely to use their hands than children who observed the hands-free demonstration, suggesting that infants imitate rationally according to the goal or intent of the demonstrator.

Along these lines, Scott and Baillargeon (2013) found that sixteen-month-old infants looked longer when actors performed less efficient actions. An actor was to remove one of two identical objects from one of two transparent boxes—one with a transparent lid, and one with no lid—sixteen-month-old infants looked longer when the actor removed the object from the box with the lid. Infants also looked less when the actor removed an object from the box with the lid if the actor had communicated a preference for the object in the box with the lid (she had always chosen that object in pretest trials). These results also suggest that infants are surprised when informants act inefficiently towards or inconsistently with their goals.

Topal, Gergely, Miklosi, Erdohegyi, and Csibra (2008) attribute A-not-B error to the pedagogical cuing inherent in these experiments. In the A-not-B paradigm an experimenter repeatedly places an object into the same one of two containers and then, on a test trial, places the object in the opposite container (Piaget, 1999). On each trial, infants are to retrieve the object, but young infants, 10-months and younger, often fail the test trial, looking for the object in the container the object was placed in on the previous trials. The results of Topal et al. (2008) suggest that when an experimenter conducts the A-not-B task while ignoring the infant that error is reduced and that if the social context is removed entirely that the error does not occur at all, stating that in the communicative context, infants are learning about the containers, specifically that container A is for toys.

Bonawitz et al. (2011) found that pedagogical demonstrations affected children's exploration. Preschoolers were exposed to a novel toy with several hidden functions. In the pedagogical condition the informant said "Look at my toy! This is my toy. I'm going to show you how my toy works. Watch this!" then pulled a tube which caused a squeak. In the naive condition the experimenter accidentally caused a squeak by pulling the tube as she set the toy on a table. Children in the naive demonstration condition explored more and found more functions of the toy than children who were given the pedagogical demonstration. The authors suggest that children in the pedagogical condition inferred there to be fewer functions because had there been more the informant would have demonstrated them. In a similar paradigm Buchsbaum, Gopnik, Griffiths, and Shafto (2011) found that children who watched a presumably pedagogical demonstrator perform a causal sequence of actions were more likely to over-imitate than those who observed a naive demonstrator.

The above results represent a small subset of pedagogy research and support many of the tenets of Natural Pedagogy including that infants make inferences about informants' intentions and learn differently based on these inferred intentions.

Though there are many studies supportive of the theory, Natural Pedagogy is difficult to prove. The crux of Natural Pedagogy is that it is innate. Claims of innateness require evaluation of newborns. As researchers seek to evaluate increasingly complex claims on increasingly young children, they stress the explanatory power of the simple motor and looking-time tasks they employ. Pedagogy research does not begin until about one year primarily because it is difficult to study infants any younger. Behavior at one year is not indicative of an innate process. Children learn many concepts in their first year, could pedagogy not be among them?

1.3. Statistical learning

We could argue that children begin with a minimal probabilistic learning mechanism that assumes all data are sampled randomly from their true generative process. *Statistical learning* is the process through which people implicitly acquire information about distributions of elements in input (Aslin & Newport, 2012).[*] The chief medium through which people learn under statistical learning is frequency monitoring. For example, infants learn language by referencing how often specific words are or are not uttered in specific contexts. In their classic study, Saffran, Aslin, and Newport (1996) found that 8-month-old infants are able to segment words given statistical information. Infants listened to two-minute recordings of sequences of syllables in a made-up language. Syllables were played in monotone frequency and at a constant rate, thus the only information infants had to learn words was the transition probabilities of syllables (the frequencies with which certain syllables follow others). After familiarization, infants listened to short sequences of words in the made-up language and non-words (sequences of syllables that did not adhere to the transition probabilities). Infants listened longer to the non-words. Kirkham, Slemmer, and Johnson (2002) showed statistical learning in 2-month-olds in the visual domain using a similar paradigm by constructing a set of visual words from sequences of shapes. During the test, infants looked longer at novel sequences. These results suggest that infants learn transition probabilities using frequency statistics.

The literature supports statistical learning as an innate, cross-modal ability. Gervain, Macagno, Cogoi, Peña, and Mehler (2008) used imaging (near-infrared spectroscopy, NIRS) to demonstrate that newborns differentiate patterns. The temporal and left frontal areas of newborns' brains showed increased activation over trials to patterned sequences of made-up syllables than to random sequences. Bulf, Johnson, and Valenza (2011) reproduced the results of Kirkham et al. (2002) in one- to three-day-old newborns.

[*] We shall use this psychological definition (as opposed to machine learning's notion of statistical learning theory) throughout the document when referencing statistical learning. These topics will receive a more formal, mathematical treatment in chapter 3.

The primary issue with incorporating statistical learning into a developmental theory is that it does not account for social learning. That is not to say that statistical learning and pedagogical learning are mutually exclusive, they correspond to learning under different sampling assumptions.

1.4. Learning to learn socially

Csibra and Gergely (2006) motivate the idea of innate pedagogy in terms of a evolutionary "just-so story" in which pedagogy is necessary for *meditated* tool use whereby tools are used for multiple, varying tasks; not created on–the–fly from immediately available materials for some one–off task and subsequently discarded once the task is done. Imagine that an early hominid infant observes an early hominid adult use a tool to chip pieces away from a log. How does the infant make inductive generalizations about the function of the tool with no background knowledge? The goal of the action is not apparent. For example, the goal of using a tool to peel aways the flesh of a fruit may be apparent to an infant who loves fruit. One may learn through so-called "blind imitation" but according to Csibra and Gergely (2006) a blind imitator "[. . .]runs into the risk of repeating the observed action when it is not appropriate, and replicating many elements of the action that are idiosyncratic to the observed individual or situation, but are irrelevant with respect to the functional use of the tool." (p. 4). As an extreme example, picture a child who after having observed his father using a hammer to fix some derelict piece of carpentry, uses a hammer to fix a sick doll while playing hospital. This is where the just-so story breaks down. The child need not be born with mechanisms that prevent this, he or she only need acquire them before it is dangerous not to have them. For example, before they are able to perform first aid with woodworking implements.

A vital part of learning to learn socially is learning from whom to learn. There is a entire literature dedicated to how children allocate their *epistemic trust*. That is, how children decide which informants are reliable or unreliable and how children learn from these informants. Epistemic trust and pedagogy should be connected in the sense that the same social learning concepts should explain both literatures.

Statistical learning represents a minimal level of abstraction: aggregating frequencies into a summary statistic.* We must devise a way to integrate statistical learning into a developmental account of social learning by using statistical learning to build a higher-level of abstraction. We can accomplish this, in part, by exploiting teleology (see Gergely & Csibra, 2003), that is, by linking ostensive cues with the unique and frequent features of the data that follow them. The research suggests that children's differential expectations of informants' behavior can be observed at less than a year of age (Topal et al., 2008; Tummeltshammer, Wu, Sobel, & Kirkham, 2014). If infants are not born with this expectation, they appear to acquire it very early.

Our developmental story proceeds as follows: Learners *learn* a pedagogical assumption very early on (before pedagogy researchers get to them). An innate pedagogical assumption is not beneficial to a newborn surrounded by a small number of helpful, knowledgeable informants. However, to learn pedagogy very early on there must be sufficient input very early on to demonstrate the concept. In the case of pedagogy, an infant must first learn the cues to pedagogy (ostensive cues) and then must have sufficient data from informants to distinguish data sampling in pedagogical and non-pedagogical contexts. That is, the infant must have sufficient data to link ostension to ostensive cues. To the naive infant, informants will then appear separate from the world, as all–knowing and ever–helpful entities. However, as the infant gains knowledge of her own, she will begin to notice idiosyncrasies in informants' data selection behavior. She must then account for these idiosyncrasies or otherwise accept inaccurate information from unreliable informants. We posit that children respond primarily by relaxing their assumption that all informants are helpful while maintaining some flexibility in their beliefs about informants' knowledgeability.

We shall use computational methods to argue that there exists sufficient input for infants to learn pedagogical sampling (chapter 2). We shall formalize a probabilistic model of learning from informants (chapter 3) that captures more sophisticated trust behavior (chapter 5) and use it to conduct developmentally–oriented computational analyses on the epistemic trust research (chapter 4). We shall use computer simulations to investigate

*To incorporate an analogy, statistical learning is akin to the *physical stance* whereas pedagogical sampling can be thought of as the *intentional stance*(see Dennett, 1978) whether the pedagogical assumption proposed by Natural Pedagogy is beneficial to learners (chapter 6). We shall show that input exists, trust in children develops differently than predicted by Natural Pedagogy, and that innate pedagogy offers no benefit to naive learners over statistical learning.

2. A COMPUTATIONAL ARGUMENT FOR THE EARLY AVAILABILITY OF PEDAGOGICALLY OPTIMIZED DATA*

The purpose of this work, as originally written is to demonstrate that though infant–directed speech (motherese) appears irrational, that it is consistent with input optimized to *teach* phonetic categories. We display it here as evidence that infants receive input very early on that demonstrate pedagogical data selection.

2.1. Introduction

Children learn language from the statistics of their input (Saffran et al., 1996), but often the statistics of the input children receive differs markedly from the statistics of normal speech. Infant-directed speech (IDS, also known as motherese) is characterized by reduced speed, elevated pitch and affect, and unusual prosody. At a less perceptible layer, Corner vowels (/ɑ/, as in pot; /i/, as in beet; /u/, as in boot) are hyperarticulated in IDS (Kuhl et al., 1997; Cristia & Seidl, 2013; Burnham, Kitamura, & Vollmer-Conna, 2002), resulting in an increased vowel space. Hyperarticulation should improve the learnability of vowel categories; example clusters that are more distant are easier to identify. This sparked the idea that IDS is optimized to teach phonetic categories (Kuhl et al., 1997). Further research demonstrated that the trend of hyperarticulation in IDS is not constant across all phoneme pairs; some pairs are hypoarticulated (Cristia & Seidl, 2013). Additionally, the within-phoneme variability of some phoneme categories increases (de Boer & Kuhl, 2003; Cristia & Seidl, 2013; McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013). The acoustical properties of IDS appear to compete with each other. Hyperarticulation makes

*This work, "Infant-directed speech as optimal input for learning vowel categories", is under review and was conducted in part with Patrick Shafto, Naomi Feldman, and Thomas Griffiths.

clusters more distinguishable while hypoarticulation and increased variability make clusters less distinguishable.

This competition has inspired efforts to indirectly evaluate the optimality of IDS by asking whether machine learning algorithms learn better from infant-directed data than adult-directed data (de Boer & Kuhl, 2003; Kirchhoff & Schimmel, 2005; McMurray et al., 2013). However, the results are mixed. There are many candidate explanations for the differences in results, but one key limitation of these approaches is they do not directly formalize optimality and therefore do not specify the expected properties of IDS.

In this paper we broadly investigate the qualitative properties of IDS by directly modeling the optimal input for teaching phonetic categories and the effect these data have on learning. Quantitative comparisons are precluded by the current absence of adult-directed speech (ADS) and IDS data sets that are phonetically and dimensionally complete and matched. Nevertheless, we demonstrate that the contentious and counterintuitive properties of IDS are consistent with input optimized to teach infant-like language-learners normal speech.

2.2. Model

Data optimized for teaching derive from the inverse of the learning process. Rather than asking what model is most likely given the current data, we ask what data are most likely to help the learner induce the correct model. This framework leads to data that are most representative of their categories (Tenenbaum & Griffiths, 2001) in the sense that they make the correct inferences more probable, while making other, incorrect inferences less probable (Shafto & Goodman, 2008; Shafto et al., 2014).

Building on previous research formalizing phonetic categories, we adopt a Gaussian mixture model framework (de Boer & Kuhl, 2003; Vallabha, McClelland, Pons, Werker, & Amano, 2007; N. H. Feldman, Griffiths, Goldwater, & Morgan, 2013; McMurray, Aslin, & Toscano, 2009). Each phonetic category is represented as a multidimensional Gaussian in formant space. Formants are the representative frequencies of vowel phonemes and manifest as peaks in the spectral envelope. The first formant is the lowest frequency peak, the second formant is the second lowest, and so on. We focus on the first and second formants, denoted F_1 and F_2, which we capture with 2-dimensional Gaussians.

We formalize phonetic category acquisition as learning an infinite Gaussian mixture model (Rasmussen, 2000; Anderson, 1991). Under this framework the learner observes some data and infers the number of categories, where each category is a Gaussian with unknown mean and covariance matrix. Learning is a problem of inferring these parameters, Θ, for each category as well as the assignment, Z, of each example to one of an unknown number of categories given some input, X, from a speaker. According to Bayes' rule,

$$P(\Theta, Z|X) = \frac{P(X|\Theta, Z)P(\Theta)P(Z)}{P(X)}. \tag{2.1}$$

Optimal teaching data derive from the inverse of the learning process. Rather than sampling data randomly from the true distribution, optimal data for teaching are sampled from the distribution that leads learners to the correct inference (Shafto & Goodman, 2008; Shafto et al., 2014). Thus teaching involves directing learners' inferences; not just toward the correct hypothesis, but away from alternatives. This corresponds mathematically to normalization over all possible data, X,

$$\begin{aligned} P_{\text{opt}}(X|\Theta, Z) \propto & \frac{P(\Theta, Z|X)}{\int_X P(\Theta, Z|X)\mathrm{d}X} && (2.2) \\ = & \frac{P(X|\Theta, Z)P(\Theta)P(Z)}{P(X)} \Bigg/ \int_X \frac{P(X|\Theta, Z)P(\Theta)P(Z)}{P(X)}\mathrm{d}X. && (2.3) \end{aligned}$$

Data are more representative of a hypothesis to the degree that the true hypothesis has higher probability relative to alternative hypotheses (Tenenbaum & Griffiths, 2001; Shafto & Goodman, 2008; Shafto et al., 2014). The learner's inference over alternative hypotheses is captured by the marginal likelihood of the data, $P(X)$. The teacher makes inferences over alternative data choices keeping the learner's inferences in mind. The optimization of the choice of data is captured by the normalizing constant.

2.3. Methods

We focus on twelve American English vowel phonemes and their first and second formants, F_1 and F_2 (Hillenbrand, Getty, Clark, & Wheeler, 1995). The data comprise 48 examples of each phoneme from female speakers. Examples with unmeasurable formant values were discarded, leaving two phonemes, /ɑ/ and /ɔ/ (as in bought), with 47 examples. The target model was derived from the means and covariance matrices calculated from each phoneme's examples (The full list of phonemes and their means and variances can be found in Table 2.1).

The normalizing constant in Equation 2.2 (also Equation 2.25 in section 2.6) is analytically intractable. We use Metropolis sampling to sample from the distribution of optimal data without having to calculate the normalizing constant (Hastings, 1970). The Metropolis-Hastings algorithm can be applied to draw samples from a probability distribution with density $p : x \rightarrow \mathbb{R}^+$ when p can be calculated up to a constant. That is, when there exists a function $f(x)$, where $p(x) = cf(x)$ and c is a constant. A proposal distribution, $q(x'|x)$, is defined that proposes new samples, x', given the current sample, x. Beginning with a sample, x, a proposed sample, x', is drawn from q. The acceptance ratio, A, is calculated from f and q,

$$A = \frac{f(x')q(x|x')}{f(x)q(x'|x)}. \tag{2.4}$$

It is easy to see that

$$\frac{f(x')q(x|x')}{f(x)q(x'|x)} = \frac{cf(x')q(x|x')}{cf(x)q(x'|x)} = \frac{p(x')q(x|x')}{p(x)q(x'|x)}. \tag{2.5}$$

If q is symmetric, that is $q(x'|x) = q(x|x')$ for all x, x', then $\frac{q(x|x')}{q(x'|x)}$ (the Hastings ratio) cancels from the equation, leaving,

$$A = \frac{f(x')}{f(x)}, \tag{2.6}$$

from which we calculate the probability with which x' is accepted,

$$P(x'|x) = \min\left[1, A\right]. \tag{2.7}$$

To sample from the distribution of optimal data using the Metropolis algorithm, we calculate the numerator exactly via enumeration and propose symmetric Gaussian perturbations to resample data. The acceptance probability is thus

$$P(X'|X) = \min\left[1, \frac{P(X'|\Theta, Z)P(X)}{P(X|\Theta, Z)P(X')}\right]. \tag{2.8}$$

For the simulations, the sampler simulated one datapoint for each phoneme (twelve total). X comprised twelve data points, one for each phoneme. X was initialized by sampling data from the prior (see section 2.6). At each iteration, new data, X', were generated from X by adding Gaussian noise distributed $N(0, 1225)$. This distribution was chosen so that the acceptance rate of X' was approximately 0.2. X' was then accepted according to Equation 2.8.

		mean		variance		
IPA	e.g.	F_1	F_2	F_1	F_2	F_1-F_2
æ	bat	675.69	2334.75	4800.43	25183.64	-4417.63
ɑ	pot	916.36	1525.83	8449.84	15615.80	4354.50
ɔ	bought	801.02	1188.28	5172.15	16614.68	6057.43
ɛ	bet	727.06	2062.67	5431.93	20408.23	-848.77
e	bait	535.48	2525.62	3962.77	23863.86	-1938.88
ɝ	Bert	523.96	1588.25	2049.91	15860.40	-443.44
ɪ	bit	484.31	2369.10	1181.03	22330.69	-182.84
i	beet	437.25	2761.31	1650.06	21738.56	1277.86
o	boat	555.46	1035.52	6496.21	15020.30	6953.69
ʊ	put	518.65	1228.56	1695.72	20907.53	2399.33
ʌ	but	760.19	1415.67	3312.88	13318.10	2538.87
u	boot	459.67	1105.52	1496.06	42130.34	-417.93

Table 2.1: List of phonemes in International Phonetic Alphabet transcription with means and variances calculated from the data.

2.4. Results

Figure 2.1 shows the distributions of the ADS vowels and the model predictions for IDS along the first and second formants. The model predicts that the statistically optimal data do not simply parrot the target distribution but modify it in ways that match infant-directed speech. Specifically, consistent with previous research (Kuhl et al., 1997; Cristia & Seidl, 2013; Burnham et al., 2002) the corner vowels are hyperarticulated. Figure 2.2 shows the predicted change in distance between all pairs of vowels. The model predicts hyperarticulation in most vowel pairs and consistent with IDS, but contra previous arguments (Cristia & Seidl, 2013), the statistically optimal input includes hypoarticulation of some vowel pairs. Figure 2.3 shows the predicted effects on within-category variability. Consistent with IDS (de Boer & Kuhl, 2003; Cristia & Seidl, 2013), but contra previous arguments (McMurray et al., 2013), the statistically optimal input includes increases in within-category variability for some categories.

To analyze the effects of optimized data on learning, an infinite Gaussian mixture model (IGMM) sampler (Rasmussen, 2000) was given 50 examples from each phoneme for optimal and ADS data. Optimal examples were generated according to the process outlined in the model specification; AD examples were generated from the estimated ADS distribution. The samplers were initialized from the prior. The 1000$^{\text{th}}$ sample (assignment vector, Z) was taken from 1000 independent runs of the IGMM for randomly-selected simulated data and randomly-generated AD data. Inferred assignment vectors were compared with the correct assignments vector via the adjusted Rand Index (ARI, see Hubert & Arabie, 1985).

The ARI measure takes on values from -1 to 1 with expected value 0, and assumes the value 1 when the two assignment vectors are identical. For two clusterings (assignments) $\mathbf{U}$ and $\mathbf{V}$ of N data points into i and j clusters (categories) ARI is computed as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{N}{2}}. \tag{2.9}$$

where n_{ij} is the number of datapoints assigned to i in $\mathbf{U}$ and j in $\mathbf{V}$, a_i is the sum $\sum_j n_{ij}$, and b_j is the sum $\sum_i n_{ij}$.

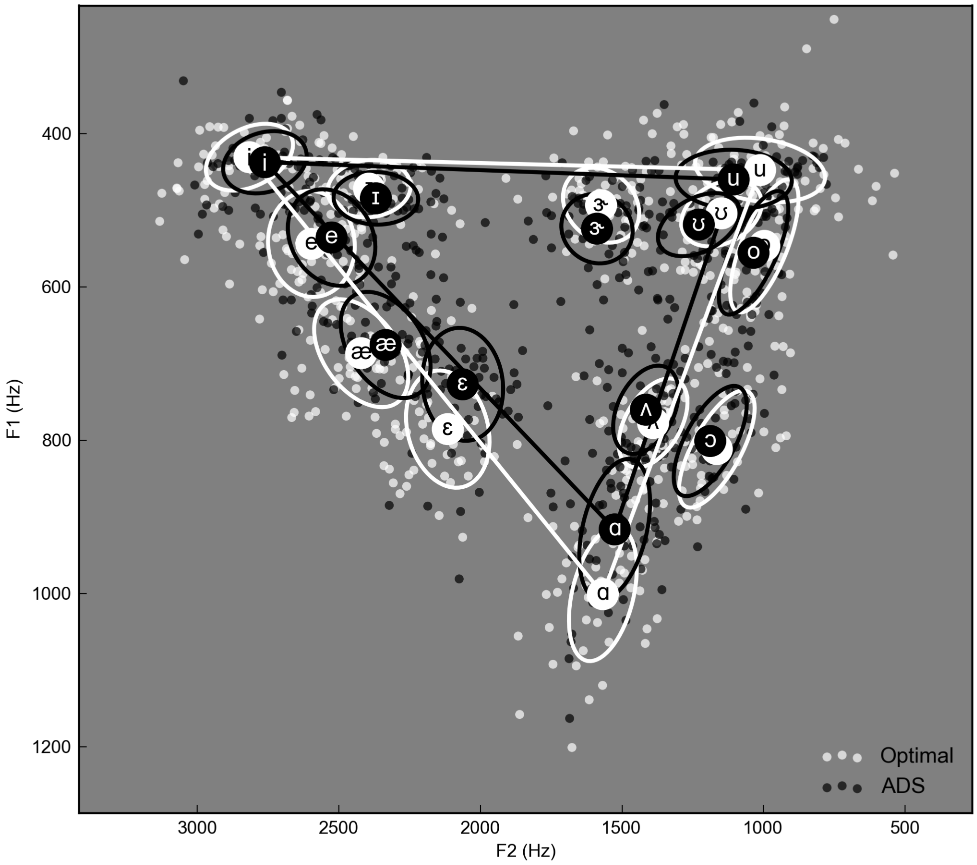


Figure 2.1: Distributions of vowels along first and second formants (F1 and F2) in adult-directed speech (black) and speech optimized for the learner (white). Differences in distributions correspond to the properties of infant-directed speech. Labels are placed at each mean, ellipses represent covariance matrices, and points are a randomly-selected subset of samples from the optimal data and the full set of adult data. The corner vowels are linked by lines. The data was generated from a single run of the sampler in which the first 1000 samples were discarded then each 100^{th} sample was collected until a total of 2000 samples had been collected.

The optimal data lead to better classification (higher ARI) in aggregate than the AD data (M_{opt}=0.3968, S_{opt}=0.0327 and M_{AD}=0.3379, S_{AD}=0.0262; d=1.9858).

Hyoparticulation can improve categorization when it is the result of disambiguating movement—that is, movement away from a cluster that it may be mistaken as a part. Increased variability can be used to mitigate any negative affects of hypoarticulation by making close or overlapping clusters more distinguishable from each other. Imagine two

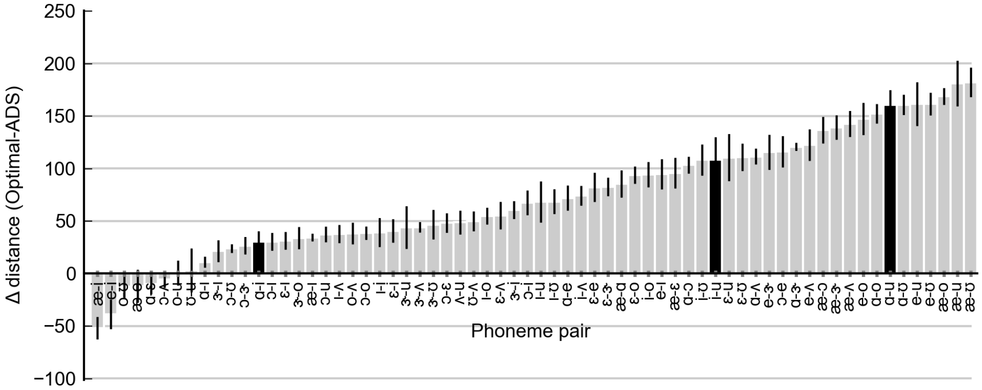


Figure 2.2: Mean change in Euclidean distance between phonemes pairs from ADS to optimal data over ten independent simulations. Black bars represent corner vowel pairs. Error bars represent standard error. For each simulation, the first 1000 samples were discarded, then each 20^{th} sample was collected until a total of 1000 samples had been collected.

very closely overlapping, circular clusters: examples from these clusters may appear to come from one large cluster. If we wish to express that there are two clusters we could stretch each cluster perpendicularly so the resulting data manifest as an 'X' rather than a single Gaussian blob.

The optimal data offer similar examples of how hypoarticulation and increased variability, when employed systematically, can be beneficial for learning. The upper left corner vowels (/i/; /ɪ/, as in bit; /e/ as in bait) remain similarly distant but increase in variance in such a way as to make individual categories more distinguishable. Specifically, /i/ significantly increases its F_1-F_2 covariance, rotating to reduce overlap. Similarly with /ɪ/. The phoneme /e/ rotates and increases its F_1 variance and emerges prominently from the others. Indeed, /i/, /ɪ/, and /e/ are better classified in the optimal data (M_{opt}=0.0895, S_{opt}=0.0399 and M_{AD}=0.0518, S_{AD}=0.0409; d=0.9315).

The distinctness of clusters is not solely a function of overlap. For example, the upper-right phonemes (/ɝ/, as in Bert; /u/; /ʊ/, as in put; /o/, as in boat) are difficult to distinguish in AD speech. In the optimal data /u/, /ʊ/, and /o/ are pressed into each other (hypoarticulated) which makes /ɝ/ more distinguishable. The corner vowel /u/ greatly increases its F_2 variance and decreases its F_1-F_2 covariance and /o/ greatly increases its F_1 variance. This causes /o/ and /u/ to overlap through each other. Their tails then emerge

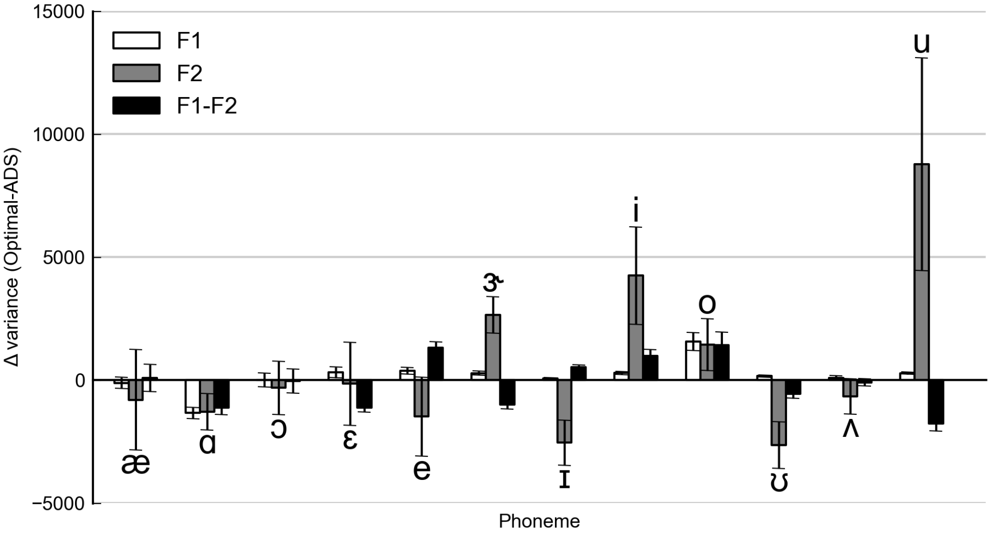


Figure 2.3: Mean change in F_1 variance, F_2 variance, and covariance ADS to optimal data over ten independent simulations. Error bars represent standard error. For each simulation, the first 1000 samples were discarded, then each 20^{th} sample was collected until a total of 1000 samples had been collected.

conspicuously from the main mass of examples which makes them more identifiable. /u/, /ʊ/, /o/, and /ɝ/ are also better classified in the optimal data (M_{opt}=0.1937, S_{opt}=0.0290 and M_{AD}=0.1296, S_{AD}=0.0790; d=1.0779).

In this way, the model captures a trade-off: the means and variances of each cluster cannot be learned if the learner has not figured out how many clusters there are. Thus, the means and variances adjust to optimize the chances of a globally correct inference. This confirms that at least some patterns of hypoarticulation and increased variability can lead to improved learning and that the mere existence of hypoarticulation or increased variability is not a conclusive argument against the optimality of IDS.

2.5. Discussion

We have defined optimal teaching data and have shown that many of the unusual and counterintuitive properties of IDS—hyper- and hypoarticulation and increased variability—are consistent with the optimal data for teaching new language learners the phonetic category model of normal, adult speech. In doing so, we have addressed many of the limitations of previous work.

First and most importantly, previous research exploring the optimality of IDS has done so without directly defining what it means for data to be optimal. We have defined optimality in terms of the data that would teach the learner the correct hypothesis, in this case, the phonetic category model of adult speech. By analyzing the optimal input for teaching adult phoneme categories, we have shown that surprising and counterintuitive properties of IDS are consistent with the optimal input for learners.

Second, existing computational studies have employed models that miss key aspects of the problem infants face. Infants do not know the number of phonetic categories beforehand (see de Boer & Kuhl, 2003), nor are they provided with feedback as to which parts of utterances belong to which categories (see McMurray et al., 2013). We have employed a computational framework that better captures the learning problem infants face: learning an unknown number of categories from unlabeled input. This, in turn, allows us to capture that input optimized to teach the phonetic category model must be optimized to teach the number of categories, as well as their locations and shapes.

Third, much of the empirical and computational research on IDS focuses on small sets of phonemes. The original work focused only on corner vowels (Kuhl et al., 1997; de Boer & Kuhl, 2003) which provide an unreasonably truncated view of the learning problem. This is made evident by more recent research which found that the trend of hyperarticulation found in the corner vowels is not constant across all phoneme pairs (Cristia & Seidl, 2013). Studying optimality in any small subset of phonemes is dangerous for it assumes that the unobserved categories do not provide added information about changes in the observed categories. Using a larger set of phonemes allows us to capture how transformations can interact in interesting ways to optimize learners' global inferences.

Just as experiments that focus on small subsets of phonemes or a small subset of learning lead to a distorted picture by oversimplifying the learning problem, intuitive theorizing can similarly lead to oversimplification. Precisely specified theories allow us to, at minimum, check our intuitions and can lead to surprising or counterintuitive predictions that intuition alone does not.

2.5.1. Limitations. The output of the teaching model depends on the model it is to teach. Though we use the most comprehensive dataset available to us (Hillenbrand et al., 1995) it does not encompass the full set of American-English phonemes and the examples for each phoneme are taken from a small set of words.[*] To achieve the most accurate representation of optimal speech data we must teach a complete representation of the phonetic category model. In order to provide a quantitative assessment of similarity between what we have defined as optimal data and IDS it would be useful to base our predictions on more accurate and complete representations of the vowels that infants learn.

2.5.2. Conclusion. Researchers have assumed that children are powerful statistical learners. This hypothesis, however, fails to account for the curious properties of language spoken to infants. Children are given odd speech, but nevertheless learn to speak normally. We have argued that the properties of IDS are consistent with input optimized to teach phonetic categories. On this view, children's ability to learn key properties of language derives not just from their powerful learning mechanisms, but also from input optimizations enacted by knowledgeable teachers. Our research suggests why parents, who intend to teach, may provide their children with unusual speech.

2.6. Detailed model specification

Here we describe the mathematical details of the model. We construct the teaching model from the learning model.

2.6.1. Learner model. We formalize phonetic category acquisition as learning an infinite Gaussian mixture model (Rasmussen, 2000; Anderson, 1991). A Gaussian mixture model comprises a set of K multidimensional Gaussian components with parameters, $\Theta = \{\theta_1, \theta_2, \dots \theta_K\}$ and an N-length vector, Z, assigning each datapoint to a component. The parameters for each component consist of a mean, μ_k and a covariance matrix, Σ_k; $\forall \theta_k \in \Theta, \theta_k = \{\mu_k, \Sigma_k\}$. Mixture model components are analogous to categories.

[*] Phonemes were measured only from words beginning with an 'h' sound e.g., /ɑ/, /i/, and /u/ were taken only from the words 'hod', 'heed', and 'who'd' respectively.

Learning is then a problem of inferring Θ and Z. Prior distributions on individual components, θ, correspond to a learner's prior beliefs about the general location (μ), and the size and shape (Σ) of categories. We assume

$$\theta \sim \text{Normal Inverse-Wishart}(\mu_0, \Lambda_0, \kappa_0, \nu_0), \tag{2.10}$$

which implies

$$\Sigma_k \sim \text{Inverse-Wishart}_{\nu_0}(\Lambda_0^{-1}), \tag{2.11}$$

$$\mu_k | \Sigma_k \sim \text{Normal}(\mu_0, \Sigma_k/\kappa_0), \tag{2.12}$$

where Λ_0 is the prior scale matrix, μ_0 is the prior mean, ν_0 is the prior degrees of freedom, and κ_0 is the number of prior observations. For simulations, we chose minimally informative prior parameters derived from the data.

$$\nu_0 = 3, \tag{2.13}$$

$$\kappa_0 = 1, \tag{2.14}$$

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{N} X_i, \tag{2.15}$$

$$\Lambda_0 = \frac{1}{K}\sum_{k=1}^{K} \Sigma_{X_k}, \tag{2.16}$$

where Σ_{X_k} is the empirical covariance matrix of the adult data belonging to category k. The prior mean, μ_0, is the mean over the entire data set, and the prior covariance matrix, Λ_0, is the average of each category's covariance matrix (see Table 2.1).

To formalize inference over the number of categories, we introduce a prior on the partitioning of data points into categories via the Chinese Restaurant Process (Teh, Jordan, Beal, & Blei, 2006), denoted CRP(α), where the parameter α affects the probability of new components. Higher α creates a higher bias toward new categories. Data points are assigned to components as follows:

$$P(Z_n = k | Z_{1 \dots n-1}, \alpha) = \begin{cases} \frac{N_k}{n-1+\alpha} & \text{if } k \in 1 \dots K \\ \frac{\alpha}{n-1+\alpha} & \text{if } k = K+1 \end{cases}, \tag{2.17}$$

where K is the current number of components and N_k is the number of data points assigned to component k. One is a minimally informative value of α corresponding to a uniform weight over components.

The standard learning problem involves recovering the true model, defined by Θ^* and Z^*, from the data, X, according to Bayes' theorem,*

$$P(\Theta^*, Z^* | X) = \frac{P(X|\Theta^*, Z^*)P(\Theta^*)P(Z^*)}{P(X)}. \tag{2.18}$$

The Normal Inverse-Wishart prior allows us to calculate the marginal likelihood, $P(X)$, analytically (Murphy, 2007), thus, for a small number of data points we can calculate the above quantity exactly via enumeration. The numerator is

$$P(X|\Theta^*, Z^*)P(\Theta^*)P(Z^*) = P(Z^*) \prod_{k=1}^{K} P(\theta_k^*) P(X_k | \theta_k^*), \tag{2.19}$$

where X_k denotes the data assigned to component k, $P(Z^*)$ is the probability of the target assignment under CRP(α), $P(\theta_k^*)$ is the prior probability of the component parameters, and $P(X_k|\theta_k^*)$ is the likelihood of the data assigned to components k given the parameters of component k.

The denominator is calculable by summing over all assignment vectors, and integrating over all component parameters

$$\begin{aligned} P(X) &= \sum_Z \int_\Theta P(X|\Theta, Z) P(\Theta) P(Z) && (2.20) \\ &= \sum_Z P(Z) \prod_{k=1}^{K_Z} P(X_k^Z), && (2.21) \end{aligned}$$

*We omit the prior distributions parameters for brevity. $P(Z)$ denotes $P(Z|\alpha)$ and $P(\Theta)$ denotes $P(\Theta|\mu_0, \Lambda_0, \kappa_0, \nu_0)$.

where K_Z is the number of components in the assignment Z, and $P(X_k^Z)$ is the marginal likelihood of the data assigned to component k in assignment Z under a Normal Inverse-Wishart prior.

2.6.2. Teacher model. Optimal data for teaching are sampled from the distribution that leads learners to the correct inference and away from incorrect inferences (Shafto & Goodman, 2008; Shafto et al., 2014). The teacher must consider the learner's inferences given all possible data, thus we normalize over all possible data X,

$$\begin{aligned} P_{\text{opt}}(X|\Theta^*, Z^*) \propto & \frac{P(\Theta^*, Z^*|X)}{\int_X P(\Theta^*, Z^*|X)\mathrm{d}X} && (2.22) \\ = & \frac{P(X|\Theta^*, Z^*)P(\Theta^*)P(Z^*)}{P(X)} \Bigg/ \int_X \frac{P(X|\Theta^*, Z^*)P(\Theta^*)P(Z^*)}{P(X)}\mathrm{d}X. && (2.23) \end{aligned}$$

The term,

$$\frac{P(X|\Theta^*, Z^*)P(\Theta^*)P(Z^*)}{P(X)}, \tag{2.24}$$

is the posterior probability of the true hypothesis given the data—the learner's inference. The learner's inference over alternative hypotheses is captured by the marginal likelihood of the data, $P(X)$. The teacher's optimization of the choice of data is captured by the normalizing constant,

$$\int_X \frac{P(X|\Theta^*, Z^*)P(\Theta^*)P(Z^*)}{P(X)}\mathrm{d}X. \tag{2.25}$$

We avoid the need to calculate this quantity directly by sampling from P_{opt} using the Metropolis algorithm (Hastings, 1970).

2.7. Summary

This research explored the significance of IDS in teaching language to infants. Infants are exposed to both adult– and infant–directed speech, though IDS is preceded by ostensive cues such as eye contact. This provides statistical learners with all the components needed to develop a pedagogical understanding: many examples of normal, random data not preceded by ostensive signals (ADS) and many examples of pedagogical data preceded by ostensive signals (IDS). It is possible that IDS does not only provide data optimized for learning speech, but also data optimized for learning about informants.

3. A PROBABILISTIC MODEL OF LEARNING FROM INFORMANTS

In this chapter, we shall use the formal language of mathematics to more exactly describe the machinery of data selection (*sampling*). Mathematical models complement empirical research by helping us to explain the phenomena we observe. For decades, psychologists and philosophers have employed mathematical models to provide *computational level* analyses (Marr, 1982) of cognitive phenomena (see Shepard, 1987; Anderson, 1991; J. Feldman, 2000).[*] We begin by formalizing sampling and learning from teachers and extend that formalization to account for data from informants who are not necessarily teachers.

3.1. Probabilistic models of data selection

We create a *probabilistic* account of data selection using Bayes' rule. Bayes' rule, named for—but (perhaps) never penned by—the Reverend Thomas Bayes (Bayes & Price, 1763), states that the posterior probability of a specific hypothesis, h_i, about the world given some data, d, is,

$$P(h_i|d) = \frac{P(d|h_i)P(h_i)}{P(d_i)}. \tag{3.1}$$

A teacher is concerned with conveying the hypothesis. A teacher who samples given only the constraint that data should be produced in proportion with their probability under the target hypothesis chooses data in proportion to the likelihood function: $P(d|h_i)$. Data that are inconsistent with the hypothesis have zero probability under the likelihood and hence will not be generated. This is the same notion of random sampling that we outlined above and found not to align with our assumptions about teaching. In spirit with the definition

[*] Computational level analysis of a cognitive system refers to an account, decoupled from the way the problem is represented in the brain, of what the system does to solve a problem.

given in section 1.3 we refer to learners who adopt this sampling assumption as *statistical learners*. The statistical learner infers the hypothesis not singly on the likelihood but also by weighting it with her current belief (prior probability) about the truth, $P(h_i)$. This means that given identical likelihood values, the learner will attribute lower posterior probability to *a priori* less likely hypotheses.

If we wish to account for pedagogical data selection, we must account for learners' expectations and inferences about teachers. Teaching is orthogonal to the machine learning concept of supervision in which a computational learner is provided a training set of labeled data from which to infer a function that maps new examples to labels. A teacher may or may not produce labels—a teacher may label an object but also may silently demonstrate its function. We have seen that (adult) human learners expect teachers to produce specific, non–random data. Learners expect teachers to choose data to induce in them a specific hypothesis. Teachers choose data knowing that learners expect them to choose data in this way. The way teachers teach and the way learners learn are intertwined and recursively defined. The probability with which a teacher will produce specific data given a hypothesis is proportional to the probability that a learner will learn that hypothesis from the data. The probability that a learner will learn a hypothesis given some data is proportional to the probability that the informant will produce that data given that hypothesis, multiplied by the baseline probability of that hypothesis. We capture *pedagogical sampling* with the following recursive equations (Shafto & Goodman, 2008; Shafto et al., 2014):

$$P_{\text{learner}}(h_i|d) \propto P_{\text{informant}}(d|h_i)P(h_i), \tag{3.2}$$

$$P_{\text{informant}}(d|h_i) \propto P_{\text{learner}}(h_i|d), \tag{3.3}$$

which capture the fact that teachers choose the data that not only best demonstrate the true hypothesis, h_i, but maximize the probability for the learner of the true hypothesis, h_i, over all alternatives. We seek to understand what assumptions learners make about others' selection of data and how these assumptions change with experience.

3.2. Epistemic trust model specification

The models of data selection discussed in section 3.1 focus on a specific type of informant: informants who choose data to cooperatively convey the true state of the world to learners. Informants of this type—pedagogs—must be knowledgeable about the world and choose data helpfully. Informants are not always knowledgeable, nor do they always have the skill or intention to teach. People are often unknowledgeable or unhelpful, and may have motivations that compel them to conceal information from learners. As we have seen previously, people learn differently from people than they do from the world. Similarly, people learn differently from informants with different features. To list a few of the documented phenomenon, people prefer informants who: label common objects accurately (Pasquini, Corriveau, Koenig, & Harris, 2007; Koenig & Harris, 2005; Fitneva & Dunfield, 2010), produce more understandable errors (Einav & E. J. Robinson, 2010; Kondrad & Jaswal, 2012), and are part of an agreeing majority (Corriveau, Fusaro, & Harris, 2009; Chen, Corriveau, & Harris, 2012). Here we outline a probabilistic model that learns from and about informants of differing epistemic qualities which we shall then employ to explain empirical findings.

We explain learners' inferences about informants as a function of inferences about informants' knowledgeability and helpfulness (Eaves & Shafto, 2012; Shafto, Eaves, Navarro, & Perfors, 2012). A trustworthy informant must posses accurate knowledge about the world and must be willing and able to share that knowledge effectively. For example, a knowledgeable but unhelpful informant may be perceived as deceptive or, less malevolently, unskilled. Unknowledgeable but helpful informants will perform actions that are informative, but because they may hold inaccurate beliefs, not necessarily maximally so. For example, any action taken on a causal system is informative because the system will react (produce an effect) which provides information about the system. An unhelpful informant may choose not to act.

The model is represented as a Bayesian Network (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993): a set of variables (nodes) causally linked by probabilistic relationships (edges). Edges link parent nodes to their child nodes. In Figure 3.1a we see a graphical

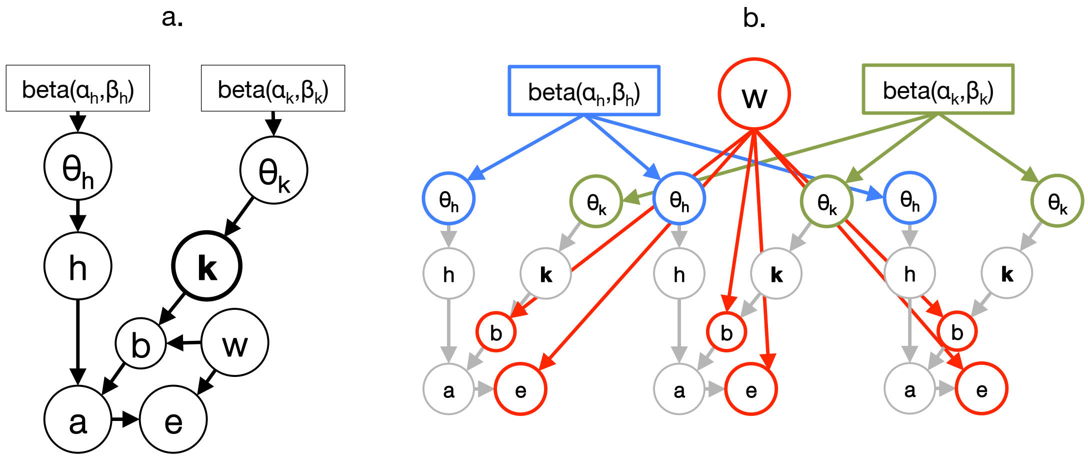


Figure 3.1: Graphical representation of the model. Informants' beliefs, b about the world, w, are determined by their knowledgeability, k, about the world. Informants' actions, a, are determined by their beliefs and their helpfulness, h. Actions on the world result in effects, e. θ_k and θ_h represent individual informants' probability of being knowledgeable and helpful, respectively. θ_k and θ_h have beta prior distributions. a) single-informant model. b) Multi-informant or group demonstration representation for a single demonstration from three informants. Note that beta priors on knowledgeability and helpfulness and the true state of the world, w, are shared across informants. Arrows and nodes are colored-coded for clarity.

representation of the model. The model represents a learner's model of how informants choose data. Informants' beliefs, b about the world, w, are determined by their knowledgeability, k, about the world. Knowledgeable informants' beliefs align with the true state of the world; unknowledgeable informants' beliefs are determined randomly according to some potentially domain– and experience–specific probability distribution. An unknowledgeable informant's beliefs may follow a uniform distribution corresponding to a completely random guess or may follow a distribution that allows some beliefs to be less likely. For example, given the true label of an animal, *lion*, an informant should be less likely to guess the label *car* than to guess the label *tiger*.[*]

Informants' actions, a, are determined by their beliefs and their helpfulness, h. Here we employ the pedagogical sampling model discussed in section 3.1. Helpful informants act to induce their own beliefs in learners; unhelpful informants act to induce other beliefs in learners. This is captured by the recursive equations:

[*] We shall explore this idea further in subsection 5.2.7

$$\begin{aligned} P_{\text{learner}}(b|a) &\propto P_{\text{informant}}(a|b)P(b), && (3.4) \\ P_{\text{informant}}(a|b) &\propto (P_{\text{learner}}(b|a))^{\alpha}. && (3.5) \end{aligned}$$

These are Equations 1.2 and 1.3 from section 3.1. We have replaced the hypothesis with beliefs and data with action. The informant cannot condition her actions on the true state of the world, only what she believes to be true because what informants know—what they believe to be true—is subject to the effects of their knowledgeabilty. Action has replaced data because the learner observes the action; the action is the data. The effect, e, does not appear in this equation. Informants must consider all the possible effects of their actions, thus effects are marginalized (summed) out of the equation. The exponent, α, controls helpfulness. When α is 1, informants choose pedagogically to produce data that maximizes the learners' belief in the belief the informant holds. When α is 0 informants choose randomly (recall $x^0 = 1$). When α is negative, informants choose deceptively, that is, to lead learners away from their belief (with some caveats as we shall see in the following section). Equation 3.4 corresponds to a shared meta-reasoning between learners and informants as discussed in section 3.1.

Actions on the world result in effects e. The effect is determined by the particular causal structure and action. In word learning, we typically disregard the effect, for unless the speaker is a wizard or has uttered some extraordinarily breathy statement, words do not elicit observable effects from the world.

Prior distributions are placed on informants' helpfulness and knowledgeability, corresponding to learners beliefs about individual informants and informants in general:

$$\begin{aligned} h|\theta_h &\sim \text{Bernoulli}(\theta_h) && (3.6) \\ \theta_h &\sim \text{beta}(\alpha_h, \beta_h) && (3.7) \end{aligned}$$

and similarly for knowledgeability,

$$k|\theta_k \quad \sim \text{Bernoulli}(\theta_k) \tag{3.8}$$

$$\theta_k \quad \sim \text{beta}(\alpha_k, \beta_k). \tag{3.9}$$

The value of h and k are determined by flips of θ-weighted coins. θ_k and θ_h are drawn from beta distributions. The beta distribution is a continuous probability distribution on the interval $(0, 1)$ defined by two parameters, α and β. We use the standard beta distribution parameterization, beta(α, β), which distributes probability according the function

$$f(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \tag{3.10}$$

where $\mathrm{B}(\cdot, \cdot)$ is the beta function. The mean of the beta distribution is $\frac{\alpha}{\alpha+\beta}$. If $\alpha > \beta$ more probability lies toward 1, if $\alpha < \beta$ more probability lies toward 0, and if $\alpha = \beta$ equal probability mass lies on both sides of .5 (see Figure 3.2). These beta distributions leave the model with four free parameters: α_k, β_k, α_h, and β_h.

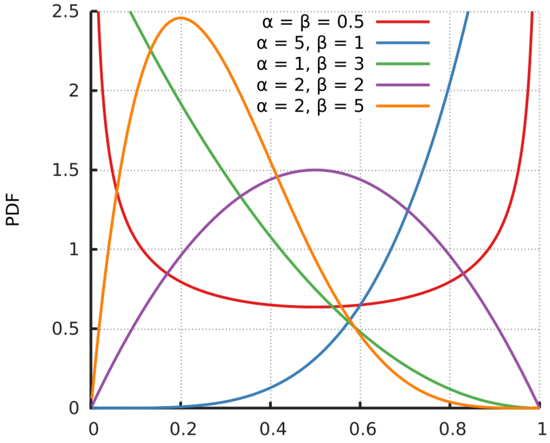


Figure 3.2: Beta distributions for a variety of parameter values. The x-axis represents the value of the beta random variate, $x \in (0, 1)$. The y-axis represents the value of the beta probability density function (PDF) at x. Image adapted from Wikipedia under creative commons license.

Beta distributions represent the distribution of people and each informant is a draw of θ from that distribution (see Figure 3.1b). Theta values persist across multiple demonstrations by a single informant. Keeping these rules in mind, we can link several single-informant graphs by their beta priors and by the state of the world to form a group demonstration (see Figure 3.1b). We can also link a number of single informant graphs by θ_k and θ_h to form successive demonstrations from a single informant. For multiple demonstrations, we need not (necessarily) link the state of the world; the state of the world is free to change from time point to time point. We can similarly link graphs in both ways simultaneously to form successive group demonstrations.

Before we proceed to apply the model to capturing empirical results, it is important to evaluate some of the assumptions we have made with respect to how unhelpful informants behave.

3.3. The difficulty of deception or: *never go in against a Sicilian when death is on the line*

Imagine: you are a locked in a game of wits with a masked man who intends to kidnap what you have rightfully stolen (Goldman, 2007). On the table before you are two goblets of wine—one of which the masked man has poisoned. You must employ all of your cunning and use what you know about your opponent to divine the location of the poison, after which you must choose a cup and drink. First, and most importantly, you know that it is the masked man's intention to mislead you—he wishes to conceal the location of the poisoned cup. You begin to reason. Only a great fool would place the cup in front of his opponent, so clearly you cannot drink from the cup in front of the masked man. But, your opponent must know that *you* are not a great fool, so you clearly cannot choose the cup in front of you. After much recursive reasoning, you have failed to reach a fixed point. Your reasoning has not settled sufficiently to create a confident inference. In fact, averaged over iterations of reasoning, each cup is equally likely to have been poisoned. You choose a cup randomly and drink. You have been tricked! Both cups were poisoned and you die—a shame. We see that doing and responding to deception without other information to leverage (say, effects from the world) are difficult tasks indeed.

Our previous work (Eaves & Shafto, 2012; Shafto et al., 2012) utilizes a set of assumptions: that a helpful informant will always utter the label he believes to be correct and that an unhelpful informant will never utter the label he believes to be correct (he will utter any word other than the correct label). If we assign different probabilities to labels given cues, with the correct label being most likely, then given a helpful informant we reach a fixed point in line with those previous assumptions. The matter is more complicated when deceiving. In our recursive model, deceptive intent is represented mathematically as a negative exponent. At each iteration of reasoning the exponent inverts the probability of producing a particular label given a cue—just as in our game of wits. The informant deceives, but we know that he deceives, but he knows that we know he deceives, and so on. Deceptive inference under this model can achieve a stable fixed point—for example in causal learning where inferences are marginalized over possible effects—but in word learning, they do not. We are left with two choices—apart from solving paradoxes which have plagued philosophers since at least 600 BC—which correspond to the different ways a deceptive informant may apply optimizations.

First, if the informant intends to teach the least likely, or most wrong hypothesis, $p(a|b, deceptive) \propto p(b|a)^{-1}$, we can invert the probabilities, $P(b|a)$ before pedagogical optimization (recursion) begins and proceed with a positive exponent, $(|\alpha|)$. Otherwise, if the informant wishes only to teach any incorrect hypothesis, we could begin with a positive exponent and then take the complement, that is, after the recursion, take $P(b|a, deceptive) = 1 - P(b|a, helpful)$. This corresponded (serendipitously) to our previous assumption that a deceptive informant will never utter the label she believes to be correct and will choose randomly among incorrect labels. It should be noted that both strategies have similar fixed points. When $|\alpha| \approx 1$ they both correspond to uttering the least likely label most often and uttering the correct label least often. However, we adopt the current strategy, that a deceptive informant wishes to teach any incorrect hypothesis,

$$P_T(a|b) \propto \begin{cases} (1 - P_L(b|a))^{-\alpha}, & \text{if } \alpha < 0 \\ P_L(b|a)^{\alpha}, & \text{otherwise} \end{cases} . \tag{3.11}$$

In any case, it is an empirical question as to which of these representations capture human behavior. No study has yet addressed this question. Here, we make a convenient assumption that aligns with previous successful endeavors.

3.4. An example from word learning

Epistemic trust studies generally follow a similar setup. Children are introduced to one or more informants from whom they receive differing data (experience) in *familiarization* trials. Children must then choose to accept or reject information from the informant(s). For example, a child may be introduced to two informants and then observe that one informant labels common objects incorrectly while the other labels them correctly (*accuracy* trials). The child may then be presented with a novel object and asked which informant he or she would like to ask for the object's label (an *ask* trial), or similarly after having observed both informants label, the child may then be asked to label the object (an *endorse* trial). Here we discuss the process by which we model these studies.

To begin, we must make some assumptions about the world. We arbitrarily assume that at any given labeling trial there are four reasonable labels.[*] That is, $|W| = 4$ and hence there are four possible beliefs, $|B| = 4$. There is a possible belief associated with each possible state of the world. In word learning, each action is a label and so the number of actions (labels) is equivalent to the number of world states and number of possible beliefs $|A| = |W| = |B| = 4$. We assume that the states of the world are distributed with uniform probability. No word is *a priori* more likely than any other

$$P(W) = \frac{1}{|W|}. \tag{3.12}$$

These assumptions result in the following relationship between the world and informants' knowledgeability and beliefs: knowledgeable informants' beliefs match the true label, w, while naive informants guess at random, uniformly from among the possible labels. The probability that an informant's belief aligns with the true state of the world is

[*] We have explored the effect of increasing and decreasing the number words and found quantitative but not qualitative differences in the model output.

$$P(b = w|k) = \begin{cases} 1, & \text{if } k = \text{knowledgeable} \\ 1/\left|W\right|, & \text{otherwise} \end{cases}. \tag{3.13}$$

As for which labels informants utter, helpful informants shall always utter the label they believe to be correct and unhelpful informants shall always utter a label they believed not to be correct:

$$P(a|h, b) = \begin{cases} 1, & \text{if } a = b \text{ and } h = \text{helpful} \\ \frac{1}{|W|-1}, & \text{if } a \neq b \text{ and } h = \text{unhelpful} \\ 0, & \text{otherwise} \end{cases} \cdot \tag{3.14}$$

Again, we focus on actions (i.e. ignore effects) in word learning demonstrations.

In familiarization trials, the model must leverage what it knows about the world to learn about informants. In accuracy trials, informants label common objects, thus the child and hence the model, knows the true state of the world. The model can then estimate the probability with which the informant is helpful and knowledgeable.* This means learning the joint probability distribution for θ_k and θ_h given a and w,

$$p(\theta_k, \theta_h|a, w). \tag{3.15}$$

During test (ask and endorse) trials, the model must use what it has learned about the informant to learn about the world. Ask and endorse questions may seem superficially similar, but they are in fact deeply dissimilar. Framed in a probabilistic context, the endorse problem is to determine the probability of each informant's label being correct given what is known about about informants in general (prior parameters) and past experience, ξ, with informant, i:

*Inference in the model is performed using standard approximation methods such as rejection sampling and Gibbs sampling. For details see Appendix A.

$$\begin{aligned} P(endorse_i) &\propto \sum_w P(w = a|a, \alpha, \beta, \xi) & (3.16) \\ &= \sum_{w,h,k} \iint_\theta P(w = a|a, h, k)P(h, k|\theta)P(\theta|\alpha, \beta, \xi)d\theta, & (3.17) \end{aligned}$$

where, for notational simplicity, we collapse analogous variables and parameters such that $\theta = \{\theta_k, \theta_h\}$, $\alpha = \{\alpha_k, \alpha_h\}$, and $\beta = \{\beta_k, \beta_h\}$.

The probability of endorsing informant 1 over informant 2 is,

$$P(endorse_{1,2}) = \frac{P(endorse_1)}{P(endorse_1) + P(endorse_2)}. \tag{3.18}$$

It is less obvious how to formalize the ask question. The question again is "who would you like to choose for information." There are two potential strategies. The naive strategy is to choose the informant who is most likely to label correctly in the future. That is,

$$P(ask) \propto \sum_{w,a} P(w = a|a, \alpha, \beta, \xi). \tag{3.19}$$

The more sophisticated strategy is to ask the informant from whom one is more likely to learn the correct label. This means performing some amount of meta-reasoning. The learner must infer what actions the learner would perform given every possible state of the world and then infer the probability of her learning the true state of the world given the informant's action.

$$\begin{aligned} P(ask) &\propto \sum_{b_L,w,a} P(b_L = w|a, \alpha, \beta, \xi) & (3.20) \\ &= \sum_{b_L,w,a} P(a|w, \alpha, \beta, \xi)P(w|a, \alpha, \beta, \xi)\delta_{b_L,w}. & (3.21) \end{aligned}$$

Where b_L is the learner's belief, and $\delta_{b_L,w}$ is the Kronecker delta function which assumes the value one if $b_L = w$ and the value zero otherwise. The implications of these two formalizations are significantly different under certain circumstances. There may be times when choosing to ask the most accurate informant may lead to a poorer learning outcomes. An unknowledgeable informant may be more likely than a deceptive informant to label correctly, but the deceptive informant provides indirect information about the true label by never uttering it. If the number of alternative labels/hypotheses is sufficiently small (say $|W| = 2$), an informant who never produces the correct label is as good for learning as an informant who always produces the correct label. This hypothetical scenario is an excellent example of how the kind of rigorous formalization required by computational models emphasizes gaps in our understanding. In any case, for simplicity's sake we avoid modeling the ask question where we can and where we cannot adopt the naive assumption that children choose to ask informant who are more likely to label correctly (see Equation 3.19).

4. PARAMETERIZING DEVELOPMENTAL TRENDS IN THE LITERATURE USING COMPUTATIONAL ANALYSIS

The usefulness of a computational model should extend beyond reproducing single experiments; we would like to be able to use models in a broad way. Of course, we strive to explain developmental phenomena but tasks cannot remain constant over development. Thus, across experiments, the ages of the participants are not the only variable changing. To account for age effects we are forced to account for the effects of other experimental features such as communication mode, paradigm, cultures, etcetera. This poses a challenging problem. Our model does not explicitly account for these factors: there is no 'age' node, nor 'communication mode' node in our graphical model. Indeed, it is not typical that these even be considered explanatory variables. Instead we use altered prior beliefs about informants' knowledgeability and helpfulness to account for age effects. In our previous research, we found that a model with a very strongly biased prior toward assuming people are helpful—to the point of being immutable—better explained three-year-olds' performance on the tasks we investigated (Shafto et al., 2012). To avoid accounting for a large number of effects we chose fairly homogeneous studies; there was no apparent reason that any factor in those experiments, apart from age, should require different model assumptions at an explanatory level. This left us with four studies to model.[*]

In this section we introduce a powerful method for conducting computational analyses using cross-categorization (Shafto, Kemp, Mansinghka, & Tenenbaum, 2011). The procedure allows us to infer a joint probability distribution over all relevant experimental features and model parameter values, which in turn allows us to build conditional distributions over

[*] Statistical factors such as sample size do not affect our ability to model a study because we compute the likelihood of any given experimental result under the model. The Bayesian approach to hypothesis testing is to compare the likelihoods of data under alternative models. In this way, the model does not depend on the limiting behavior of sample means (i.e. the central limit theorem) as do standard null-hypothesis models.

parameters and features. From these conditional distributions we gain the ability to ask and answer fundamental questions about how demographic features such as task and age are related to the variables in the model, e.g., how is pointing reflected in children's beliefs about helpfulness compared to verbal testimony, or how do children's beliefs about informants' helpfulness change from age 18 months to 3 years to 4.5 years. We first briefly discuss cross-categorization and then illustrate how it can be used to conduct a computational analysis by applying it to our model of epistemic trust.

4.1. Cross-categorization

Cross-categorization is a hierarchical, Bayesian non-parametric method for categorizing tabular data (Shafto et al., 2011). It groups features (columns) into *views* and within each view it groups objects (rows) into *categories.* In this section, we shall informally construct cross-categorization from the Infinite Mixture Model (IMM; see Teh et al., 2006; Rasmussen, 2000; MacEachern & Müller, 1998; Neal, 2000; Anderson, 1991, for more information on IMMs). Those looking for a detailed treatment of cross-categorization are referred to Shafto et al. (2011).

To begin, an IMM is a mixture model that places a particular generative prior on the assignment of data to components that allows for an infinite number of mixture components. A common prior process used in IMMs is the so-called Chinese Restaurant Process (CRP; see Aldous, 1985). The CRP assigns items to an existing component with probability proportional to the number of items currently assigned to that component and assigns items to new components with probability proportional to a concentration parameter, α. This leads to a *rich-get-richer* assignment scheme under which components with more items receive more items in the future and a higher value of α leads to more components. More formally, given a partition, $Z = \{z_1, z_2, \dots, z_n\}$, of n items into K components, the probability of assigning item $i \in \{1, ..., n\}$ to component $z_i = k$ is,

$$P(z_i = k | Z_{-i}, \alpha) = \begin{cases} \frac{n_{-i,k}}{n-1+\alpha} \text{ if } k \in \{1 \dots K\} \\ \\ \frac{\alpha}{n-1+\alpha} \text{ otherwise} \end{cases} , \tag{4.1}$$

where Z_{-i} is the partition Z less z_i and $n_{-i,k}$ is the number of items assigned to component k less item i.

Cross-categorization behaves as a hierarchical IMM in which features' assignments to views and objects' assignments to categories within views are each given a CRP prior. The data in each feature are modeled via a data-appropriate statistical model (usually conjugate), $P(x|\theta)$, where θ represent the distribution parameters; and a prior on θ, $P(\theta|\phi)$, where ϕ are the hyperparameters for the distribution of θ. For example, continuous data may be modeled by a Normal distribution with mean and precision parameters, $\theta = \{\mu, \rho\}$ which are in turn modeled with conjugate Normal-Gamma prior with prior hyperparameters, $\phi = \{m, r, s, \nu\}$ (Murphy, 2007; Fink, 1997). Inferences are made over ϕ, the CRP α parameter for features' assignments to views and the CRP α parameter for objects' assignments to categories within views, using some vague, empirically-derived hyper-prior. This produces a very broad model suited to a wide variety of data analysis tasks. Each cross-categorization state (or sample) provides us with a features (columns) into views, for each view a partition of objects (rows) into categories, a set of hyperparameters for feature models, and view and category CRP α parameters.

4.2. Method

The meta-analysis procedure begins by modeling a number of experimental studies; finding the parameters that cause the model output to best replicate (fit) experimental data; and building a tabular dataset from these parameter sets along with experimental demographic information of interest, where each modeled unit corresponds to a row, and the demographic variables and inferred model parameters correspond to columns. We then run cross-categorization and conduct analyses using the resulting cross-categorization samples. The idea is to use cross-categorization to create a full joint probability distribution over the data from which we can derive the conditional distributions between demographic variables (e.g. age) and model parameters. This way, we can ask targeted questions using Bayesian inference about how parameters change given certain variables while disregarding other variables.

We modeled 11 studies which we divided into analysis units. For example, an experiment which separately reported results for three- and four-year-olds consists of two analysis units. We then searched for model parameters that best reproduced the data in each unit.* To construct the cross-categorization table, we took the five† parameter sets with the lowest errors for each unit and arranged them in a table where each row represented a single parameter set for an analysis unit and each column was a parameter or a demographic or experimental feature of interest (see Table 4.1). We did not include error in the table because we did not want our ability to capture trends in the parameters to be affected by the model's baseline ability to fit an experiment. We were interested in trends in the parameter space where the best fits occur. Additionally, though we included the experiment identifiers in the table, they were ignored during analysis. We wanted to see which units grouped together but did not want cross-categorization to cluster explicitly by unit ID. We included the demographic variables that characterized the 11 experiments: mean age, experiment paradigm (e.g. looking time, ask-endorse), communication mode (e.g. verbal, pointing), and the dominant culture of participants (e.g. first-generation Asian-American immigrants, Western/American/Caucasian).

For easy interpretation, we converted the model's α and β parameters on knowledgeability and helpfulness to *strength* and *balance* (Kemp, Perfors, & Tenenbaum, 2007). The strength (s) and balance (b) parameterization of the beta distribution is

$$\theta \sim \text{beta}(sb, s(1-b)). \tag{4.2}$$

Thus,

$$\begin{aligned} s &= \alpha + \beta & (4.3) \\ b &= \frac{\alpha}{\alpha + \beta}. & (4.4) \end{aligned}$$

[*] For details regarding the studies we modeled and the parameter search procedure see chapter 5.

[†] We took the top five parameter sets to buffer the noise in the parameters that results from calculating experimental error from approximation output (recall that we use Gibbs sampling to simulate experiments in the model).

Balance corresponds exactly to the mean of the beta distribution and strength roughly corresponds to the invariance or peakedness of the distribution. A higher b_k results in a higher bias toward knowledgeability; a higher s_k results in a belief that is more resistant to updating given new data.

s_k	b_k	s_h	b_h	exp index	age	culture	comm mode	paradigm
12.8	0.5	3.0	0.9	1	3.5	Asian	verbal	ask-endorse
1.3	0.72	23.1	0.9	4	0.8	American	gaze	looking-time
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 4.1: Example structure of a hypothetical prepared table used during cross-categorization. Columns correspond to strength and balance parameters for knowledgeability and helpfulness, experiment-age identifier, age in years, communication mode, and experimental paradigm.

For cross-categorization, each feature must be assigned an appropriate probability distribution. All zero-bounded continuous features (s_k, s_h, and age), were assigned Lognormal likelihood functions with the standard conjugate Normal-Gamma prior; balance parameters were assigned Normal likelihood functions with Normal-Gamma prior; and the categorical variables (communication mode, paradigm, and culture) were assigned Multinomial likelihood functions with the conjugate symmetric Dirichlet prior. We used a custom python implementation of cross-categorization.* We collected 64 samples after 500 iterations of inference. That is, we initialized 64 independent Markov chains of the sampler, ran the sampler for 500 iterations, and conducted analyses using the 64 independent states.

4.3. Analysis and results

First, we examine the dependence probability matrix (see Figure 4.1). Each entry, (i, j) of the dependence probability matrix represents the proportion of samples in which features (columns) i and j have been clustered into the same view. That is, the proportion of samples in which feature i and j are statistically dependent. By looking at each demographic feature's dependence probability with the model parameters we can get a rough sense of which variables influence model behavior. As a reference point, the expected dependence

*Our implementation, *BaxCat*, can be found at https://github.com/BaxterEaves/BaxCat probability derived from the CRP with parameter α where $\alpha \sim \mathrm{Exp}(1)$ is roughly .596 (for the full derivation of this quantity see Appendix B).

In Figure 4.1 we see that in general the dependence probability between columns is high. Of note, we see that the strength parameter for helpfulness and age are highly dependent and that both helpfulness parameters are highly dependent with communication mode.

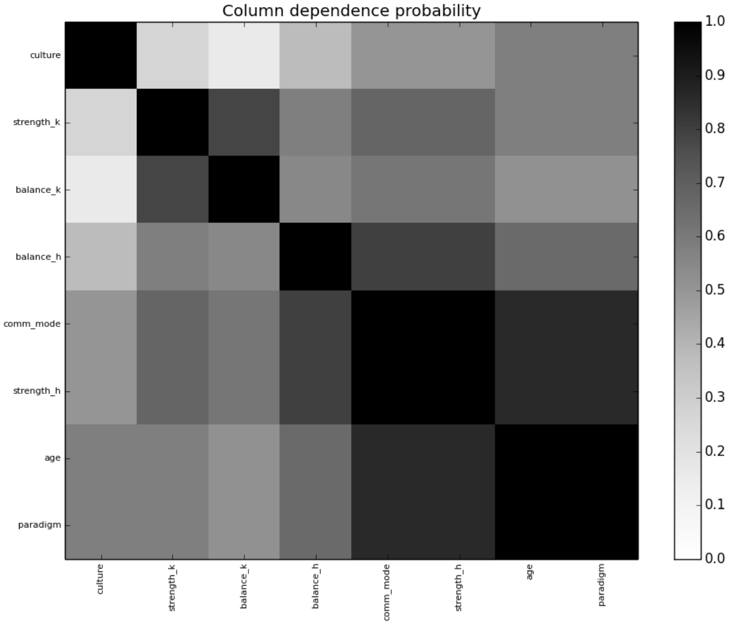


Figure 4.1: Resulting dependence probability matrix from cross-categorization. Each cell, $[i, j]$ of the table represents the probability of dependence between columns i and j. Probability is represented by shade. The lighter the shade, the lower the dependence probability.

In a single cross-categorization sample, the conditional probability of a value, x, in column i given a value, y, in column j is

$$P(x|y) = \begin{cases} \sum_{c \in C_v} \frac{n_c}{n+1+\alpha_v} P(x|X_c)P(y|Y_c) + \frac{\alpha_v}{n+1+\alpha_v} P(x)P(y) & \text{if } z_i = z_j \\ \sum_{c \in C_v} \frac{n_c}{n+1+\alpha_v} P(x|X_c) + \frac{\alpha_v}{n+1+\alpha_v} P(x) & \text{if } z_i \neq z_j. \end{cases} \tag{4.5}$$

Where n is the number of objects in the table, C_v is the set of categories belonging to view v, α_v is the CRP concentration parameter for view v, n_c is the number of objects assigned to category c, and X_c and Y_c are the data in X and Y assigned to category c. Note that if $z_i \neq z_j$—columns i and j are not in the same view—then $P(x|y) = P(x)$ because columns

i and j are independent. For conditional distributions over multiple models, we employ model-averaging. Conditional distributions are averaged over samples:

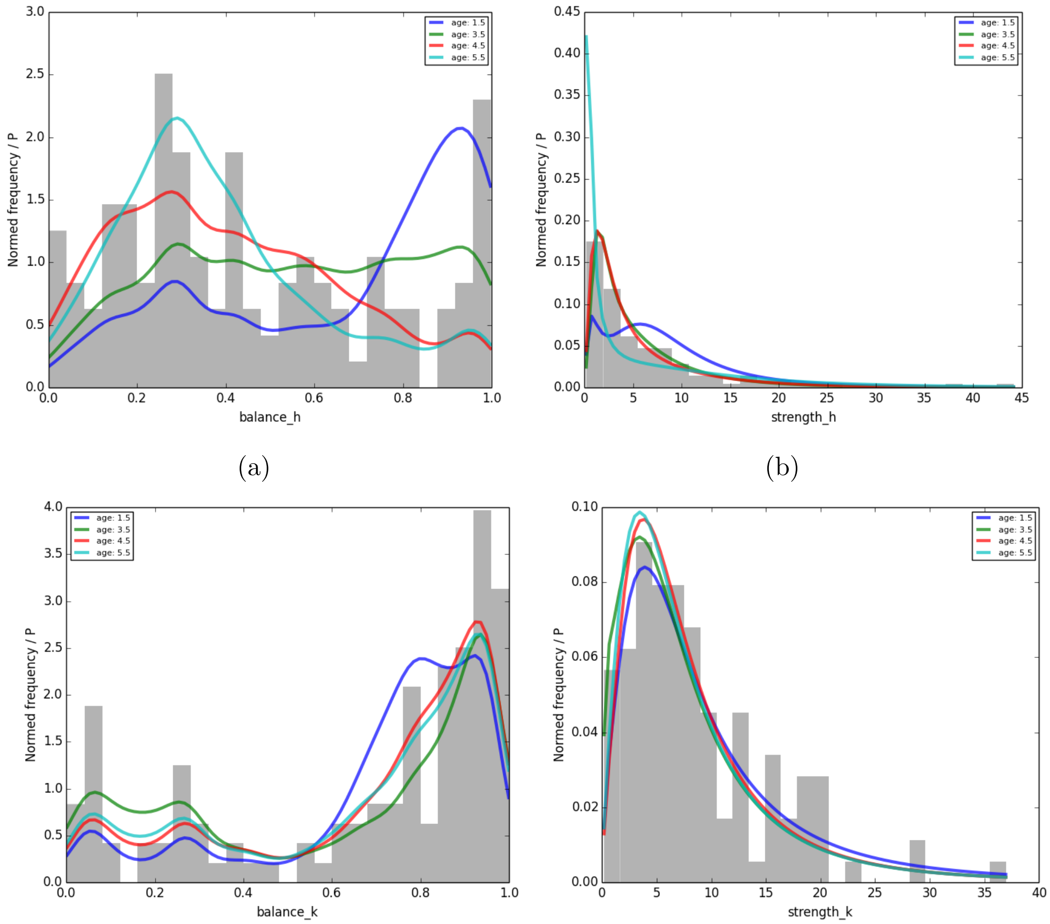
$$P(x|y) = \frac{1}{|S|} \sum_{s \in S} P_s(x|y). \tag{4.6}$$

Where S is the set of samples, s is an individual sample, and $P_s(x|y)$ is the conditional probability of x given y under sample s. We may now query the conditional distributions.

First, we wish to re-evaluate our previous findings regarding age: younger children's behavior is better explained by a model highly biased to assume that people are helpful (Shafto et al., 2012). In our previous research we evaluated this hypothesis by comparing the likelihoods of three-year-olds' experimental results given both the full model and a knowledgeability-only model. Here, we have not modified the model in any way; we simply observe what trends occur naturally. We compute the distributions for knowledgeability's strength and balance given a set of age groups, $P(s_k|age = \{1.5, 3.5, 4.5, 5.5\})$ and $P(b_k|age = \{1.5, 3.5, 4.5, 5.5\})$. We do the same for helpfulness. These distributions can be seen in Figure 4.2 (a,b).

The results confirm our previous findings. We see that the mass of balance for helpfulness for 18-month-olds rests heavily toward 1 (see Figure 4.2a), and similarly—though to a lesser extent—at 3.5 years. A more neutral balance allows for more flexible inference over a variety of informant types. When balance lies to one side of .5, the model is biased to explain outcomes in terms of that bias. A model heavily biased to believe that informants are helpful and knowledgeable has trouble explaining beliefs in terms of deception or ignorance. Balance at 4.5 and 5.5 years peaks at a more neutral position. We see a similar trend with strength. Younger ages have higher mean strengths. Higher strength values result in slower updating in the presence of new data. Thus, the model captures younger children's behavior by attributing higher, more rigid prior biases toward helpfulness. We calculated similar distributions for knowledgeabilty's parameters but saw no significant age differences (see Figure 4.2 c and d). There seems to be slightly more mass toward 1 for 18-month-olds' knowledgeability balance but the shapes of the distributions for each age group are largely identical.

We found that communication mode had a profound effect on the helpfulness balance distribution (Figure 4.3). The conditional distribution given that the informant communicated by placing makers (e.g. Couillard & Woodward, 1999) showed a markedly different trend than the conditional distributions given verbal testimony, points, or gaze. The model accounted for marker placement with a more neutral—almost non-informative—distribution. We find this interesting because of the communications modes we model, marker placement is the only one which is not explicitly an ostensive, or pedagogical cue (see Gergely et al., 2007; Csibra & Gergely, 2009). It should be noted that we modeled only a single experiment that used non–ostensive cues (Couillard & Woodward, 1999); more data are needed before we may make confident claims about how the model accounts for non-ostensive cuing.



(a)

(b)

(c)

(d)

Figure 4.2: Conditional probability distribution of helpfulness and knowledgeability parameters given age. Conditional probability distributions for the ages 1.5 (blue), 3.5 (green), 4.5 (red), and 5.5 (teal) superimposed on the normalized histogram of the original data in gray. (a) Averaged conditional probability distribution of helpfulness's balance parameter. (b) Averaged conditional probability distribution of helpfulness's strength parameter. (c) Averaged conditional probability distribution of knowledgeability's balance parameter. (d) Averaged conditional probability distribution of knowledgeability's strength parameter.

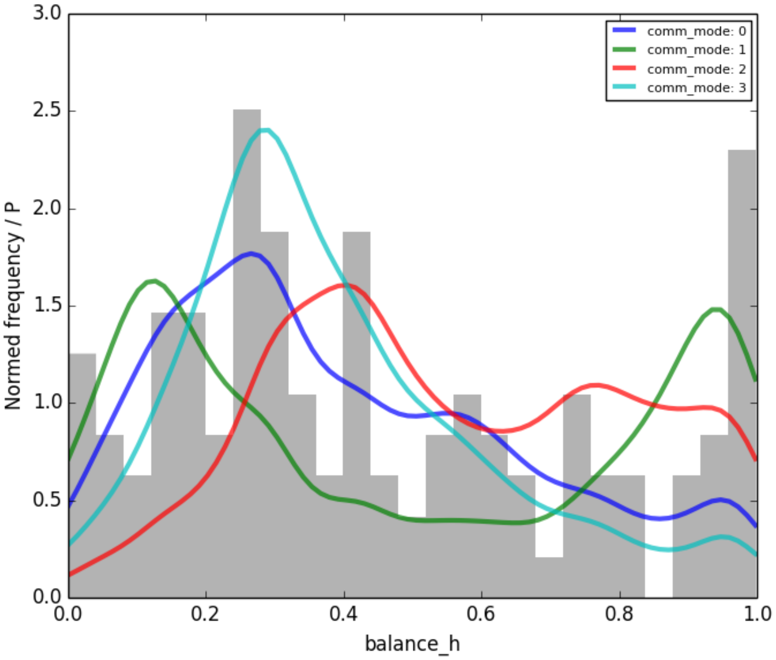


Figure 4.3: Averaged conditional distribution of helpfulness's balance distribution given each communication mode overlayed on the data (gray bars). In blue: verbal, in green: marker placement, in red: pointing; in teal: gaze.

5. MODELING EPISTEMIC TRUST ACROSS TASKS AND AGES

Now that we have formalized our model we may apply it to empirical results. The trust in testimony literature has proceeded without the guidance of any unifying theory and as a result, is scattered. There is very little overlap or reproduction between studies. In this chapter we seek to demonstrate the power of our model as a guiding theory by modeling a broad set of empirical results.

In Table 5.1 we have listed the studies we shall model along with information pertaining to the ages of the children studied, the experimental paradigm employed, the mode through which informants offered testimony, and the information-seeking strategy investigated in each study. The primary inclusion criterion for a study is that there is an easy mapping from experiment to model. That is, we have sufficient information to apply a specific mathematical formalization to a paradigm. For example, we can account for the effects of a learner's familiarity with an informant in a purely Bayesian sense (see subsection 5.2.3), but it is not clear how to account for the effect of an informant's gender. We also choose studies that look at similar strategies. Many of the studies we use focus on accuracy and consensus of and among informants but with small adjustments, which allows us to ensure the model is generalizable, and to tune the model to multiple studies simultaneously (as we shall see in section 5.1).

The goal for this chapter is to demonstrate that the model captures results that it should and does not capture results that it should not; that it is powerful but fails appropriately.

5.1. A note on parameter fitting

Before we begin it is important to discuss how the model accounts for performance difference across ages, cultures, communication modes, etc. There is no node for age, nor

Study	Strategy	Ages (years)	Comm mode	Paradigm
Koenig and Harris (2005)	accuracy	3, 4	verbal	ask-endorse
Fitneva and Dunfield (2010)	accuracy	4, 7, 19+	verbal/print	ask-endorse
Pasquini, Corriveau, Koenig, and Harris (2007)	relative accuracy	3, 4	verbal	ask-endorse
Corriveau and Harris (2009)	familiarity	3, 4, 5	verbal	ask-endorse
Corriveau, Fusaro, and Harris (2009)	consensus	3, 5	pointing	ask-endorse
Couillard and Woodward (1999)	points v. markers	3.25, 3.75, 4.25	pointing, marking	endorse
Koenig and Echols (2003)	accuracy	1.5	verbal	looking time
Chen, Corriveau, and Harris (2012)	consensus	4, 6	pointing	ask-endorse
DiYanni, Corriveau, Nasrini, Kurkul, and Nini (under review)	consensus v culture	4	verbal	endorse
Tummeltshammer, Wu, Sobel, and Kirkham (2014)	reliable gaze	.75	gaze	looking time
Einav and E. J. Robinson (2010)	error magnitude	4-5, 6-7	verbal	ask-endorse

Table 5.1: List of studies modeled. Columns correspond to the publication, the selection strategy investigated, the age groups of the children studied, the communication mode through which informants offered testimony, and the experimental paradigm.

culture, nor communication mode; the model captures these effects with different prior parameters. For example, two age groups may perform differently on an identical task; the model captures this difference with a different parameter set (α_k, β_k, α_h, and β_k) for each age group. Adjusting parameters to better replicate (*fit*) data is referred to as parameter or model fitting. One should not necessarily expect a single parameter set to account for all the data given such a heterogeneous set. Indeed, differences among the best fitting parameters, for example, across ages, can be important signs of development. Fitting the model to each experiment is dangerous as well, as a fit to a single experiment may not generalize to other experiments (over-fitting). The reasonable middle ground is to fit the model to like experiments. For example, we may group experiments into categories based on age and paradigm—experiments are grouped together if there is a valid reason to expect that they should be explained by similar parameters sets (or if they are explained by different parameter sets, as in chapter 4). Research indicates that different age groups perform differently (Pasquini et al., 2007; Koenig & Harris, 2005), so we should cluster by age; research also indicates that children respond differently to pointing than they do to marker placement (Couillard & Woodward, 1999), so we should cluster by communication mode; and so on. Table 5.2 shows our best efforts to reduce degrees of freedom while maintaining the theoretical boundaries indicated by the literature. We began with 24 different experiments/ages/conditions to reproduce and have reduced the effective number of parameter sets to 13. We then search for parameters to reduce the error across a cluster of experiments. Each results figure will represent the model output given the parameter sets determined by clustered fitting unless otherwise noted.

Study	Age (y)	Comm mode	α_k	β_k	α_h	β_h
Pasquini, Corriveau, Koenig, and Harris (2007)	3	verbal	0.048	0.641	0.777	0.677
Koenig and Harris (2005)	3	verbal				
Corriveau and Harris (2009)	3	verbal				
Pasquini, Corriveau, Koenig, and Harris (2007)	4	verbal	2.811	0.970	0.323	0.121
Koenig and Echols (2003)	4	verbal				
Corriveau and Harris (2009)	4	verbal				
Fitneva and Dunfield (2010)	4	verbal				
Corriveau, Fusaro, and Harris (2009)	3	points	7.283	0.263	8.078	0.0205
Couillard and Woodward (1999)	3	points				
Corriveau, Fusaro, and Harris (2009)	4	points	7.640	0.288	3.861	1.308
Couillard and Woodward (1999)	4	points				
Chen, Corriveau, and Harris (2012)	4	points				
Chen, Corriveau, and Harris (2012)	6	points	0.617	1.797	1.162	0.737
DiYanni, Corriveau, Nasrini, Kurkul, and Nini (under review)	5 (c)	verbal	5.197	0.343	0.275	0.696
DiYanni, Corriveau, Nasrini, Kurkul, and Nini (under review)	5 (a)	verbal				
Corriveau and Harris (2009)	4	verbal				
Couillard and Woodward (1999)	3	markers	10.692	2.232	1.341	0.068
Couillard and Woodward (1999)	4	markers	5.948	0.349	1.705	12.193
Einav and E. J. Robinson (2010)	6-7	verbal	3.067	0.120	0.888	5.791
Fitneva and Dunfield (2010)	7	verbal				
Koenig and Echols (2003)	1.5	verbal	3.388	0.865	10.003	0.282
Fitneva and Dunfield (2010)	19-22	verbal	1.644	0.260	0.0725	0.097
Tummeltshammer, Wu, Sobel, and Kirkham (2014)	.75	gaze	13.712	1.610	0.139	0.384
Einav and E. J. Robinson (2010)	4-5	verbal	0.559	8.553	0.010	9.893

Table 5.2: Clusters of experiments used for grid search parameter optimization. Grouped experiments are separated by lines. Columns list the study, the age in years of the participants, the communication mode, and the optimal parameters. DiYanni et al. (in review) appears twice in the same cluster because they look at effect of culture. (c) represent Caucasian children and (a) represents Asian children.

To achieve a fit, we employ a simple grid search procedure which involves generating a large number of parameter sets, running the model for each experiment for each parameter set, and calculating the errors. Rather than using a predetermined grid of parameter values, we generated 4000 parameter sets from independent exponential distributions with mean 5. That is, for each parameter set,

$$\alpha_k \sim \text{Exp}\left(\frac{1}{5}\right), \tag{5.1}$$

$$\beta_k \sim \text{Exp}\left(\frac{1}{5}\right), \tag{5.2}$$

$$\alpha_h \sim \text{Exp}\left(\frac{1}{5}\right), \tag{5.3}$$

$$\beta_h \sim \text{Exp}\left(\frac{1}{5}\right). \tag{5.4}$$

We choose the set of parameters that minimizes the sum relative error across experiments in each cluster. The relative error of two values, a and $b \neq 0$ is the absolute value of one minus their ratio, $|1 - a/b|$. If a/b is 1 then $a = b$. We subtract 1 from this quantity because 1 is the point that represents zero difference between a and b. We take the absolute value because we are not concerned with the direction of the error, only its magnitude. The sum relative error between two n-length vectors of values $\mathbf{a}$ and $\mathbf{b}$ is then,

$$\sum_{i=1}^{n} |1 - \mathbf{a}_i/\mathbf{b}_i| \,. \tag{5.5}$$

We use relative error rather than squared or absolute error because experiments' dependent measures and hence the model's output is not always identically scaled. For example, one experiment may report the proportion of children who asked a particular informant for information while another may report the number of seconds an infant looked at an event. We use the sum of error so that the error of each data point (bar in a bar chart) carries equal weight. An experiment with more bars should be weighted higher for error minimization.

5.2. Procedure and results

We order presentation of the experiments such that we progress from the simplest in terms of the model and then add more complicated extensions later.

5.2.1. Accuracy. One of the studies that spurred the trust-in-testimony gold rush was Koenig and Harris (2005)'s study on children's preference to ask for and endorse information from accurate sources. For three trials children observed two informants label common objects, e.g., a ball and a cup. One informant labeled each object correctly and the other labeled each object incorrectly. After these *accuracy* or *familiarization* trials, a novel object was placed before the informants. The child was either invited to choose the informant whom she would like to ask for the label (*ask* trial) or after having observed each informant provide his own label, the child was invited to label the object herself (*endorse* trials).

This study maps easily to a Bayesian inference. We have only to account for data that does or does not match the state of the world. Participants observed novel informants, thus there is no need to account a prior bias that one informant should be more likely than the other to label correctly. Additionally, each informant's incorrect answers are equally incorrect (labeling a ball as a shoe is just as foolish as labeling a cup as a dog) therefore there is no need to account for the relative *magnitude* of errors (which we account for in a later section). Endorse questions are modeled as in section 3.4.

During accuracy trials, children learn about their informants. The model is concerned with learning the probability distribution defining each informant's tendency toward or away from helpfulness and knowledgeability given the state of the world (the object) and the label uttered by the informant. This means collecting information about k and h given w and a.

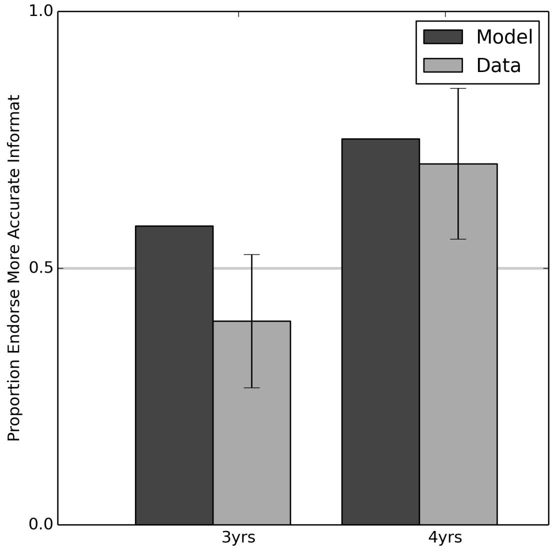


Figure 5.1: Model simulation results for Koenig and Harris (2005). The y-axis represents the proportion of children who endorsed the answer given by the accurate informant, or for the model, the probability of endorsing the accurate informant.

We see the model results along side the experimental results (Koenig & Harris, 2005, Experiment 1) in Figure 5.1. For both age groups, the model prefers to endorse the label provided by the accurate speaker. The clustered fitting has caused some loss in the model's ability to account for three-year-olds' data due to the significant variability between studies. For example Koenig and Harris (2005) and Pasquini et al. (2007, 100% v 0% condition) performed nearly identical procedures on three-year-olds and achieved markedly different results. The model should not be expected to account for different results from identical experimental paradigms using the same parameter set. The model is closer to chance (.5) for three-year-olds. In the original study, three-year-olds' preference was not significantly different from chance but four-year-olds' preference was. The model infers that an informant who always labels accurately is likely knowledgeable and helpful and that an informant who always labels inaccurately is not. In fact, an informant who repeatedly labels incorrectly is assumed to be knowledgeable and unhelpful—deceptive—because an unknowledgeable informant, regardless of whether he is helpful, will produce the correct label occasionally. An unknowledgeable and helpful informant will produce the correct label by correctly guessing; an unknowledgeable and unhelpful informant will produce the correct label upon incorrectly guessing the state of the world and producing labels to lead the learner away from that incorrect belief.

This preference for more accurate informants has been documented after even a single encounter (Fitneva & Dunfield, 2010). In Fitneva and Dunfield (2010) children were shown an image and told a corresponding story. A sticky note occluded part of each image. The child asked two informants (children on a computer screen) what was under the card. The two informants answered differently. The sticky note was removed, revealing that one informant had been correct and the other had been incorrect. The procedure was then repeated but the child was allowed to ask only one informant. For this study we modeled ask questions. The results, averaged over three experiments can be seen in Figure 5.2.[*]

We see that the model captures people's preference for the accurate informant as well as an increasing preference with age. A theme in the literature is that older individuals seem to update more rapidly given new data.

5.2.2. Relative accuracy. Informants are not deterministic. They are not always correct or always incorrect; they provide information with some amount of noise. Pasquini et al. (2007) extended the paradigm of Koenig and Harris (2005) to account for variable levels of relative accuracy between informants. Children were introduced to two informants who labeled four common objects with variable accuracy. Informants labeled either 100%, 75%, 25%, and 0% accurately, corresponding to four, three, one, and zero of four objects

[*] The procedure was identical for each experiment in Fitneva and Dunfield (2010), only the wording changed.

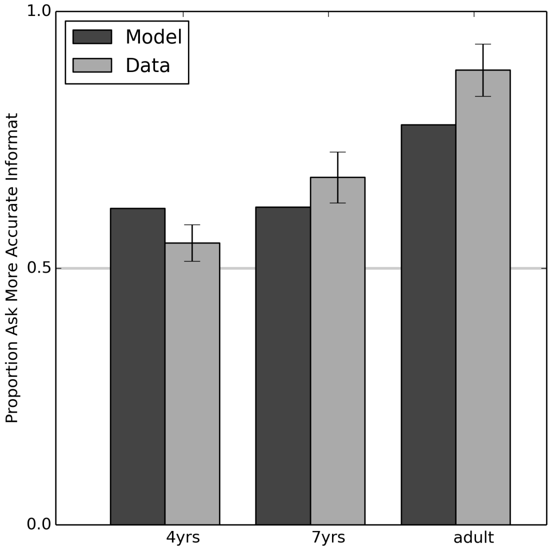


Figure 5.2: Model simulation results for Fitneva and Dunfield (2010). The y-axis represents the proportion of children who asked the previously accurate informant, or for the model, the probability of asking the accurate informant.

correctly labeled, respectively. There were four conditions 100% vs 0% accurate, 100% vs 25% accurate, 75% vs 0% accurate, and 75% vs 25% accurate. For example in the 100% vs 25% accurate condition the child observed one informant label each object correctly and the other label only one of the four objects correctly. Both informants never incorrectly labeled the same object. This feature is irrelevant to the model. The model assumes that informants label independently, and this assumption is implied in the research. The only features that affect the model's preference for an informant are the accuracy or inaccuracy of the labels and the order in which accurate and inaccurate labeling occurs. After accuracy trials, a novel object was placed before the child who then participated in ask and endorse questions.

The model (Figure 5.3) shows a preference for the more accurate informant. We see a tiered effect in both three-year-olds' behavior and model prediction. In previous research, we found that 3-year-olds behavior is best represented by a model with a strong bias toward believing all informants are helpful. This means that the model predicts three-year-olds' inferences about informants primarily on knowledgeably. Informants are either knowledgeable or not. An informant who always labels correctly is knowledgeable, all other informants are

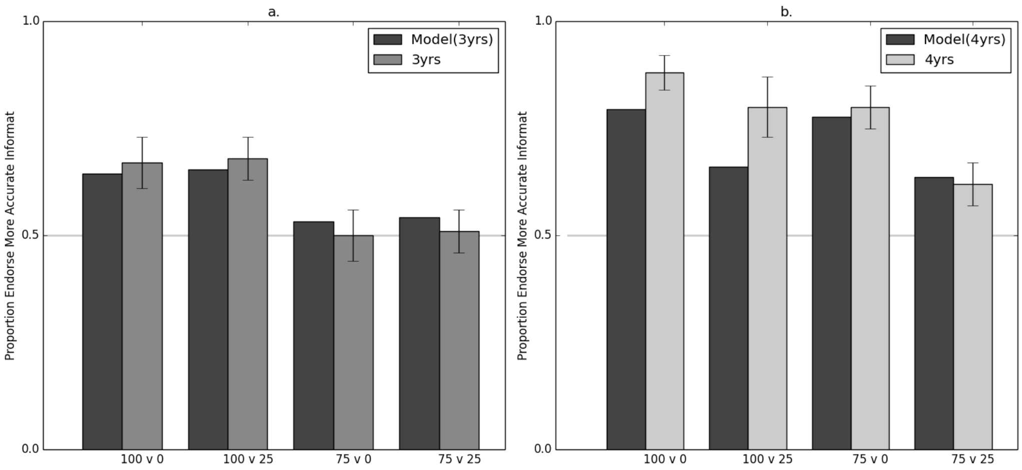


Figure 5.3: Model simulation results for Pasquini, Corriveau, Koenig, and Harris (2007) (clustered parameter fit). a) Three-year-olds. b) Four-year-olds. The y-axis represents the proportion of children who endorsed the answer given by the accurate informant, or for the model, the probability of endorsing the accurate informant. Error bars represent standard error.

not. This causes difficulty in creating a grading between the different accuracy levels.

The model predictions show a rather different trend for four-year-olds. Four-year-olds' data in the 100% vs 25% condition is not well fit. We did not see this trend in our previous work (Eaves & Shafto, 2012; Shafto et al., 2012). This mismatch is due to the clustered fitting of parameters. To establish that this is not a fundamental failure of the model, we have included the results for the single best-fitting parameter set in Figure 5.4. The best-fitting results closely follow the data, plateauing where there is a 75% difference in relative accuracy between informants. The clustered fit prefers the more accurate informant less in the 100% vs 25% accurate condition than in the 75% vs 0% condition, as observed in previous work (Eaves & Shafto, 2012; Shafto et al., 2012).

5.2.3. Familiarity. Corriveau and Harris (2009) investigated the interaction between familiarity and accuracy. For their study, Corriveau and Harris (2009) chose children's preschool teachers to play the role of familiar informants. Familiarity is formalized as prior experience. In this case specifically, because the familiar informants were teachers—not tricky uncles—we modeled familiarity as experience demonstrating helpfulness and knowledgeability. This manifests mathematically as an altered prior. This manipulation is

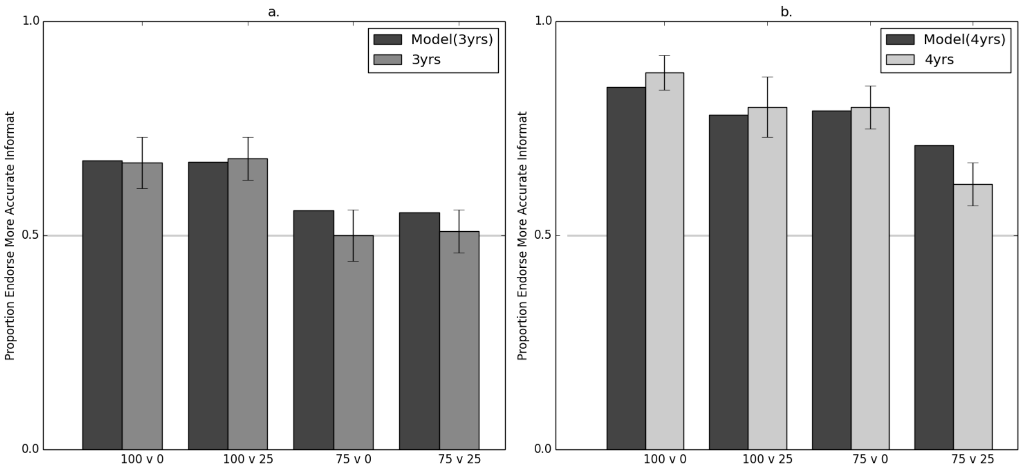


Figure 5.4: Model simulation results for Pasquini, Corriveau, Koenig, and Harris (2007) (best-fitting parameter). a) Three-year-olds. b) Four-year-olds. The y-axis represents the proportion of children who endorsed the answer given by the accurate informant, or for the model, the probability of endorsing the accurate informant. Error bars represent standard error.

straightforward to implement as a beta distribution posterior update. As a demonstration, assume that we have witnessed an informant be helpful twenty times and unhelpful once. Given a base prior of $beta(\alpha_h, \beta_h)$ the posterior distribution is simply $beta(\alpha_h + 20, \beta_h + 1)$. We used this procedure for both knowledgeability and helpfulness. The result is a strong bias and requires more data to override than the presumably weaker bias for an unfamiliar informant.

Before any familiarization or accuracy trials, children were given ask and endorse questions to gage their natural preference for the familiar informant (*pretest*). Children were then given four familiar object labeling trials in which the familiar informant labeled each object accurately and the novel informant labeled each object inaccurately (familiar 100%) or in which the converse occurred (novel 100%). If children hold a more biased belief that their teacher is helpful and knowledgeable, they should prefer to ask and endorse their teacher at pretest. Observing the teacher label common objects correctly should reinforce this bias, and observing her labeling them incorrectly should work to relax or reverse the bias.

We see in Figure 5.5 that the model captures trends across several ages but fails to capture the sharp reversal made by five-year-olds when the familiar informant labels in-

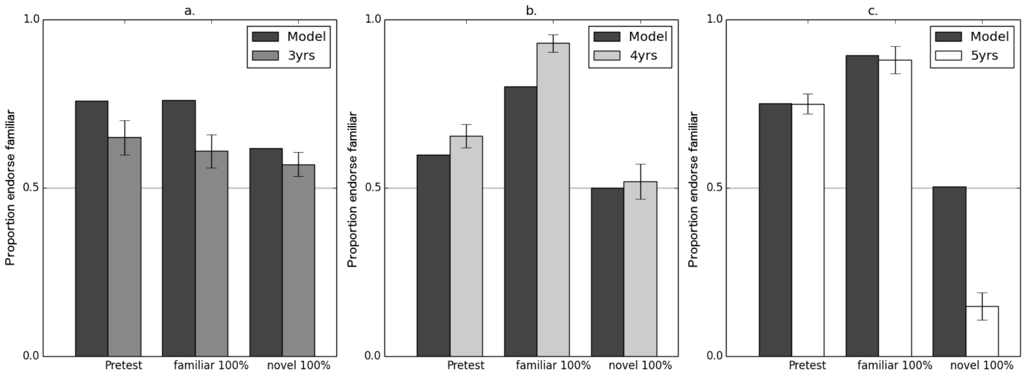


Figure 5.5: Model simulation results for Corriveau and Harris (2009). a) Three-year-olds. b) Four-year-olds. c) Five-year-olds. The y-axis represents the proportion of children who endorsed the answer given by the familiar informant, or for the model, the probability of endorsing the familiar informant. Error bars represent standard error.

accurately in the *novel 100%* condition. A possible reason for this is that to minimize complexity we have applied the same familiar prior to each age group. It is logical to assume that children of different ages have different experiences with their teachers or handle familiarity in a more flexible way (though we have not tested this hypothesis).

5.2.4. Consensus. Corriveau, Fusaro, and Harris (2009) looked at children's preferences for members of a group over rogue dissenters. For four trials, three novel objects were laid out before a group of four informants. On each trial an experimenter asked "Which is the [novel object label]", after which each informant pointed simultaneously to an object. Three informants pointed to the same object and the other pointed to a different object. On each trial the same informants agreed and the same informant dissented. It is important to emphasize that informants pointed rather than offering verbal testimony. We did not model points any differently than verbal communication. There were three objects, thus $|W| = 3$ which implies the number of beliefs, $|B|$, and number of actions, $|A|$, both equal three. After these group (pretest) trials children observed as two of the informants, one of whom had belonged to the agreeing group and the dissenter, labeled additional novel objects (test trials). Children again chose the object which they believed corresponded to the label.

We see the model results in Figure 5.6 (a and b). Because the objects were novel, children

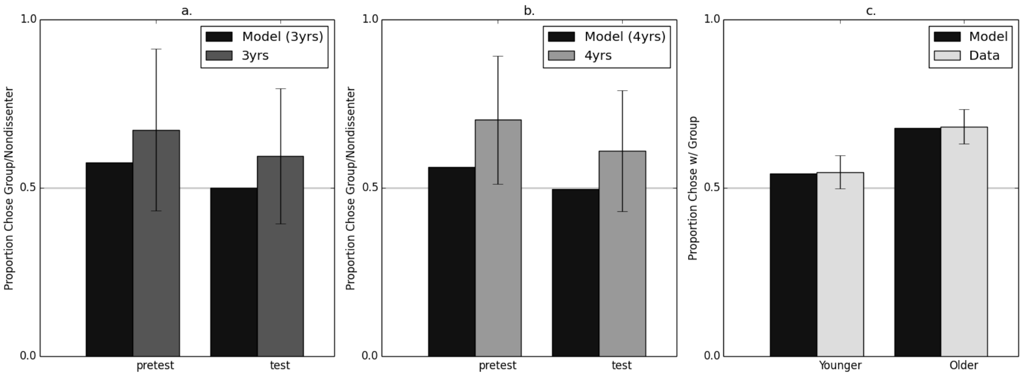


Figure 5.6: Model simulation results for Corriveau, Fusaro, and Harris (2009) and Chen, Corriveau, and Harris (2012). a) Corriveau, Fusaro, and Harris (2009), three-year-olds. b) Corriveau, Fusaro, and Harris (2009), four-year-olds. c) Chen, Corriveau, and Harris (2012) Younger and older groups.

could not leverage their knowledge of the world to learn about informants. However, the fact that children learned from a group of informants labeling the same objects provides extra power not only for learning about novel objects but learning about informants as well. In the case of a group consensus we can exploit informant dynamics. In general, it is unlikely for multiple independent informants to repeatedly converge on the same object unless they are both helpful and knowledgeable. This leads logically to the conclusion that the dissenter is either unknowledgeable, unhelpful, or both; and that the agreeing informants are pointing to the correct object.

As a simple illustration of why this is so, let us categorize informants into two groups: reliable and unreliable. Further assume that reliable informants always point to the correct object and that unreliable informants point uniformly at random. We assume that informants are reliable and unreliable with equal probability. Given three objects to choose from, the probability that three reliable informants converge on the same object is 1; the probability that three unreliable informants converge on the same object is $\binom{3}{1}\left(\frac{1}{3}\right)^3 = \frac{1}{9}$. The probability that unreliable informants converge on the same answer for four trials is then $\left(\frac{1}{9}\right)^4 = \frac{1}{6561}$.

Things are not so black and white in the model so this effect is softened. The certainty of these inferences is dependent to an extent on prior beliefs about informants. The higher the prior toward knowledegability and helpfulness, the higher the probability that agreeing informants are knowledgeable, helpful, and correct. This of course assumes uniform probability over labels. It is possible that there may be some wrong belief with a high prior probability that unknowledgeable informants could converge on.*

We also modeled the results of Chen et al. (2012) which reproduced the pretest (group) trials of Corriveau, Fusaro, and Harris (2009) with different age groups. The model procedure was identical. The results can be seen in Figure 5.6c. Again, the model captures a bias toward choosing with the group which appears to increase with age.

5.2.5. Culture.

It is not enough to demonstrate that a model fits data; the model should fail to capture results outside of its scope. Here we demonstrate how our epistemic trust model fails to account for non-epistemic, cultural behavior.

DiYanni, Corriveau, Nasrini, Kurkul, and Nini (under review) looked at culture effects in children's deferring to consensus. Children observed three informants choose a tool to crush a cookie. The tool was either functionally affordant (hard plastic) or non-affordant (a mass of plush, fuzzy balls). Each of the three informants had a cookie in front of them. The first informant selected the affordant tool and tapped the cookie twice with it then repeated the procedure with the non-affordant tool. The cookie remained intact. The informant then held the non-affordant tool and said "This is the one I would need". This process was repeated with the other two informants. Children were then asked which tool would be best for crushing the cookie. A similar condition was conducted but with a single informant. The hypothesis was that children in both culture groups would similarly reject the advice of a single informant claiming that the non-affordant tool was best, but that for cultural—not epistemic—reasons Asian-American children would be less likely to dissent from the group. For modeling purposes we treat this task as equivalent to labeling. The effect is the same in each case, the cookie remains intact, and can be ignored. Informants explicitly label the non-affordant tool as "the one I would need" which we interpret as a novel object labeling task in which one of the objects is "the best for crushing cookies". Children's bias for the affordant tool plays a major role and so we modeled the bias based on previous research using the same tools in which "[. . .]89% of 3-4-year-olds choose to use

*For example, that in the time of Christopher Columbus it was common knowledge that the Earth was flat.

the Functionally-Affordant tool over the Non-Affordant tool to crush a cookie when both tools are modeled with equal intention" (DiYanni et al., under review; DiYanni & Kelemen, 2008). The prior probability on w was left uniform because both tools are equally novel, but $P(b|\neg k, w)$ was altered such that an unknowledgeable informant should guess the affordant tool was best 89% of the time.

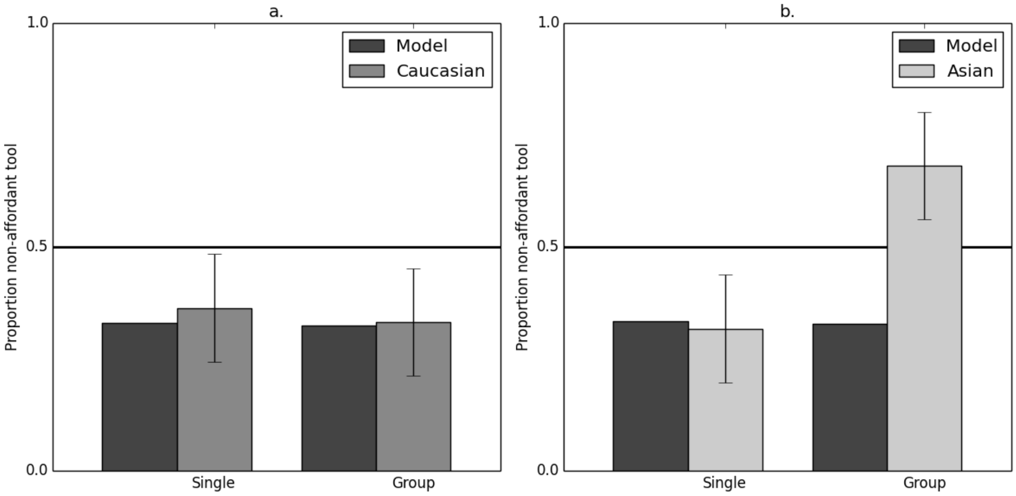


Figure 5.7: Model simulation results for DiYanni, Corriveau, Nasrini, Kurkul, and Nini (under review). a) Caucasian children. b) Asian Children. Error bars represent standard error.

Both groups of children were equally likely to dismiss the advice of a single informant, but Caucasian-American children more often rejected the advice of the consensus than did the Asian-American children. DiYanni et al. (under review) suggest that this result stems from a cultural stigma with respect to deviancy in the Asian community. The model can only venture to capture these results as modified prior beliefs (see Figure 5.7).

The model captured American-Caucasian children's disagreement with both the single informant and the group but fails to capture Asian-American children's agreement with the group. The failure is not a result of loss due to the clustered fitting; the model fundamentally fails to capture Asian-American children's behavior. The study noted that Asian-American children's conformity is likely a symptom of their avoiding appearing deviant (DiYanni et al., under review)—not an epistemic goal. In Asian cultures, dissension is stigmatized.

It is important that the model fails to capture this result because the result is non-epistemic. It is likely that group membership studies do not capture differential learning but simply the effect of social norms. Other research would suggest that children have no difficulty in appeasing a group of seemingly unreliable informants, but do not allow it to affect their learning. Corriveau and Harris (2010) demonstrated that though children may appear to defer to a group whose consensus violates their own perceptions (in the study, the group agreed a shorter line was longer than a longer line), children rely on their own perceptions when solving a pragmatic task. Though children agreed with the group that a shorter line was the longest, children then used the longest line to construct an adequate bridge to help a bunny cross a gap.

5.2.6. Deceptive pointing and marking. In Couillard and Woodward (1999)'s study on children's interpretation of deceptive points a child plays a game of Two Cup Monte with an informant. Behind a screen, the informant hides a sticker under one of two cups. The screen is taken away and the informant points to one of the cups. Children's job is to choose the cup under which the sticker is hidden, for each time children choose correctly they are allowed to keep the sticker. This procedure repeats for ten trials. On each trial the experimenter indicates the empty cup. We assume that a point acts as a label and we assume that the informant is knowledgeable because children observe the informant hide the sticker (though they do not observe under which cup). The knowledgeability bias is applied to the prior. Children receive feedback after each trial. The experiment is iterative. Each trial consists of an endorse question (choose to endorse or reject the informant's testimony) and a subsequent familiarization demonstration in which the child is given information regarding the veracity of the informant's testimony. Because the bias toward knowledgeability has been strongly influenced by the informant's hiding the sticker, children must make inferences primarily through inferences with respect to helpfulness. The informant knows the location of the sticker but does not want learners to know. Children at three-years-and-three-months of age were more often fooled by the informant than children closer to four-years of age.

The experiment was repeated with a markers condition in which the informant placed a marker to indicate a cup rather than pointing to it. Younger children were far more likely to choose the correct cup in the markers condition. We make no fundamentally different modeling assumptions to capture this result but allow it to manifest in an alternate parameter set.

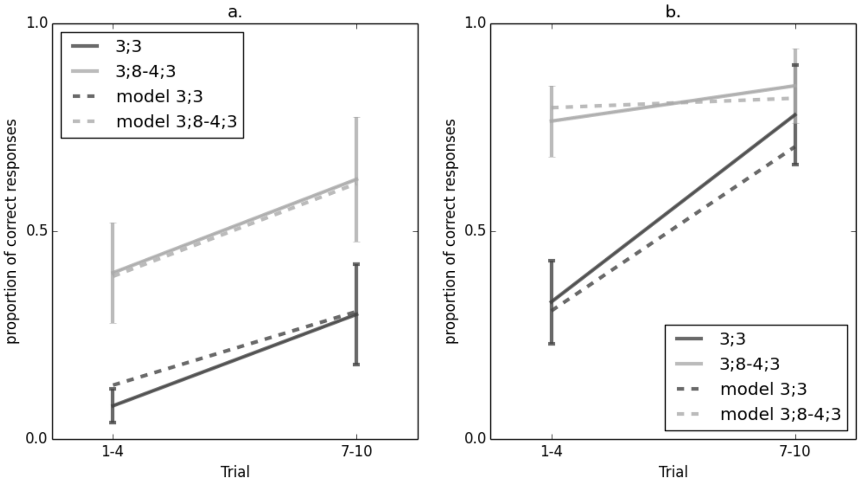


Figure 5.8: Model simulation results for Couillard and Woodward (1999). a) Points. b) Markers. The x-axis shows the trial number collapsed into blocks. The y-axis displays the proportion of children who choose the cup opposite the cup indicated by the informant, or for the model, the probability that the marker is in the cup opposite the cup indicated by the informant.

Figure 5.8 shows the proportion of children who chose the correct cup (the cup not indicated by the informant) averaged across the first four and last four trials. We see that the model captures the rate of learning. At each trial the learner is given extra information about the informant which it uses to learn about the world. The informant is reliably inaccurate. An informant who repeatedly labels incorrectly is likely deceptive. Because a deceptive informant never labels correctly, the model infers that the opposite cup is more likely. Table 5.2 shows the the primary parameter difference between three- and four-year-olds in the points task (Figure 5.8a) is in α_h. Younger children have a stronger belief that informants are helpful. A stronger belief requires more data to overcome thus we see that younger children more often choose with the informant, though they do choose with the informant less as trials progress. Table 5.2 shows that three-year-olds bias toward helpfulness has been relaxed, and that the parameter set that best captures this result for four-year-olds is biased toward deception. This strangeness in four-year-olds' parameters results in part from the way the data are collapsed into blocks. It appears that there is a slow if not non-existent learning rate across blocks and a high early preference to choose opposite the informant. It may have been that children mostly chose with the informant on the first trial and then quickly learned to choose opposite. To find parameters that best represents the model's explanation for data, it is be best to have data for each time point, but trial-by-trial data are not available. Collapsing data into block precludes the possibility of capturing the initial bias.

5.2.7. Error magnitude. Einav and E. J. Robinson (2010) looked at the effect of error magnitude on children's informant preferences. For example, labeling a lion as a tiger is a smaller magnitude error than labeling a lion as a mouse or a clock. The structure of the study was nearly identical to that of Pasquini et al. (2007). Children observed two informants label common animals for four trials. On each trial after the first, both informants labeled incorrectly but one informant produced higher magnitude errors. For example, given the labels "dog", "tiger", "horse", and "butterfly", the more accurate informant provided the labels "dog", "lion", "cow", and "bee", while the less accurate informant either provided the labels "dog", "mouse", "fish", and "cat" (animal-animal condition) or "dog", "clock", "fork", and "car" (animal-object condition).

Some words are more prevalent than others. If one was asked to provide a word starting with the letter 'A' one may be more likely to respond 'Apple' than 'Appendectomy'. To capture that some labels are more inappropriate in response to certain cues, we must formalize a meaningful relationship between words. Griffiths, Steyvers, and Firl (2007) had success using semantic networks and *Pagerank* (Page, Brin, Motwani, & Winograd, 1999; Sloman, Love, & Ahn, 1998).

The lexicon can be nicely organized into a network where associated words share links. We can represent a network containing n words as a $n \times n$ matrix $\mathbf{L}$ where $\mathbf{L}_{ij}$ is 1 if there is a link from word j to word i and 0 otherwise. Pagerank captures that important words have more incoming links and that importance travels along these links. Pagerank is thus recursively defined: important nodes have more links incoming from important nodes. If $\mathbf{M}$ is a matrix where $\mathbf{M}_{ij}$ is the total proportion of importance that travels through $\mathbf{L}_{ij}$, then

$$\mathbf{M}_{ij} = \mathbf{L}_{ij} \bigg/ \sum_{k=1}^{n} \mathbf{L}_{kj}, \tag{5.6}$$

and Pagerank is the solution for $\mathbf{r}$ in the recursive equation,

$$\mathbf{r} = \mathbf{M}\mathbf{r}. \tag{5.7}$$

Now that we have defined a prior probability distribution on cues, $p(\text{cue})$, we must define a sampling distribution (likelihood) for labels given cues, $p(\text{label}|\text{cue})$ which is exactly $P(b|\neg k, w)^*$: the probability of an unknowledgeable informant believing a particular label given the cue, w. For this we apply the idea of *spreading-activation* (Collins & Loftus, 1975) in which activation—which is directly analogous to importance—flows from node to node in the network. We can construct an activation-based sampling distribution by assuming that the probability of a label given a cue is determined by the minimal path length from the cue to the label in the network. That is, the closer the label is to a cue in a network, the higher its probability. More formally, if we assume that activation decays at the same rate across every edge, then for the set of edges, D, that defines the minimal path from *cue* to *label*, the probability of *label* given *cue* is,

$$P(\text{label}|\text{cue}) \propto \gamma^{|D|}, \tag{5.8}$$

where $|D|$ is the number of links in the path ($|D| = 0$ if $\mathit{label} = \mathit{cue}$) and $\gamma \in (0, 1]$ is a decay constant capturing that activation decreases as a function of distance. We arbitrarily chose $\gamma = .5$ which corresponds to losing half of the signal at each jump. This formalization of the belief probabilities implies that low-magnitude errors are most indicative of a helpful, unknowledgeable informant while high-magnitude errors are most indicative of unhelpful informants. A knowledgeable informant knows the correct label and an unknowledgeable informant is likely to guess a close label. In both cases, unhelpful informants will choose a label to lead learners away from their own beliefs: a label distant from the true label or distant from a close label.

[*] $\neg k$ is the negation of k or *not knowledgeable*.

The network used here was constructed from the University of South Florida free association norms database (Nelson, McEvoy, & Schreiber, 2004)[*] which comprises free associations for 5019 cue words. We only included words that were both cues and responses, leaving 4870 words. Links were created from targets to responses. We used the python package *NetworkX*[†] to construct the network, find minimal paths, and calculate Pagerank. This allowed us to model the study using the exact words used in the study rather than word analogs as in the previous studies. For example, given this model we can ask for the probability that an informant is knowledgeable and helpful given that she labeled a lion as a tiger, $P(k, h|a = \text{tiger}, w = \text{lion})$, instead of asking about a label indices, $P(k, h|a = 0, w = 1)$, or simply whether a label does not match the true state of the world, $P(k, h|a \neq w)$. It should be noted that the free-association database records responses given to text cues and not visual cues, which were used in the study.

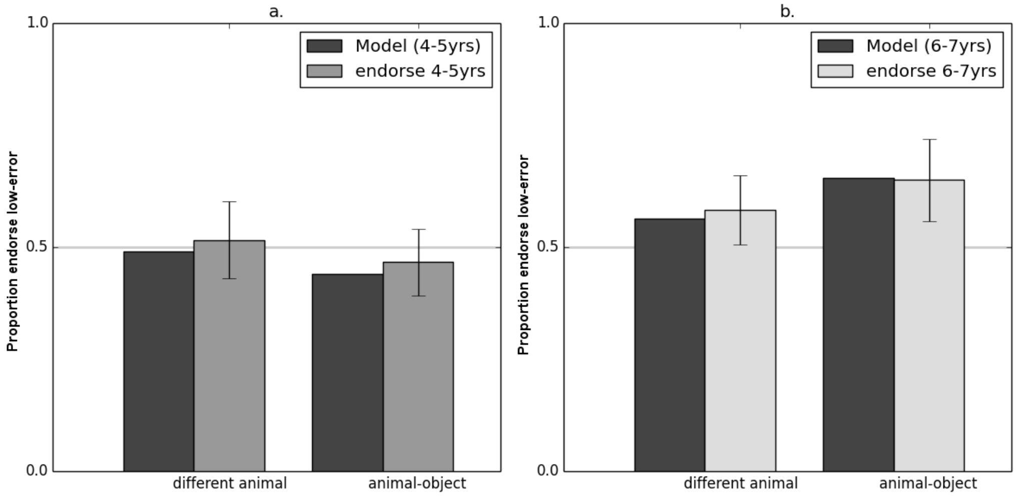


Figure 5.9: Model simulation results for Einav and E. J. Robinson (2010). a) Four- and five-year-olds. b) Six- and seven-year-olds. The x-axis displays the accuracy condition. The y-axis shows the proportion of children who endorsed the answer given by the lower-magnitude-error informant, or for the model, the probability of endorsing the lower-magnitude-error informant.

The experimental results (see Figure 5.9) indicate that four- and five-year-olds do not show a preference but six- and seven-year-olds prefer informants who produce lower-magnitude errors. Higher magnitude errors are a better indication of naivety or unhelp-

[*] Currently, the database can be found at http://w3.usf.edu/FreeAssociation/

[†] https://networkx.github.io/

fulness than lower magnitude errors. Unknowledgeable, helpful informants should guess a label close to the target and then produce a label that is close to the guessed label.

5.2.8. Looking time. The model is easily adapted to account for looking time paradigms. The primary hurdle is the mapping from probability to looking time. We assume that the time spent looking at an event is inversely proportional to the probability of that event. We are aware of recent work that suggest looking time follows a U-shaped function whereby infants look longer at moderately improbable events and less at extremely probable or improbable events (Kidd, Piantadosi, & Aslin, 2012). Recent work has successfully modeled this phenomenon (Piantadosi, Kidd, & Aslin, 2014), but adopting this model requires doubling the number of parameters in our model, which we believe adds unjustifiable complexity.

We model Koenig and Echols (2003, Study 1) in which 18-month-olds observed novel informants label common objects, displayed on a screen, either correctly (*true* labels condition) or incorrectly (*false* labels condition) for twelve trials. At each trial the number of seconds infants looked at the informant, the object, and their parents (on whose lap they sat) was recorded. We model only the time spent looking at the informant because the model most fluidly produces the probability of an informant producing a specific label given a specific target. Koenig and Echols (2003) report the mean looking time over trials. We report the mean inverse probability scaled arbitrarily. It is important to note that the parameter fit for this particular experiment was achieved by minimizing the the error of the *proportion difference* between the time spent looking at each informant in both the accurate and inaccurate conditions. For example, if infants in the true labels condition looked at the informant for an average of 4 seconds and infants in the false labels condition looked at the informant for an average of 7.5 seconds, the proportion difference is $7.5/4 = 1.875$. If the mean inverse probabilities for the true and false labels conditions are 1.2 and 3.8, respectively, then the relative error is $|1 - (3.8/1.2)/(7.5/4)| = 0.69$. We use this method because we are interested only in the trend from one condition to the other; we make no attempt to find the scaling constant that maps inverse probability to seconds. In this way, we can capture the trend without adding complexity.

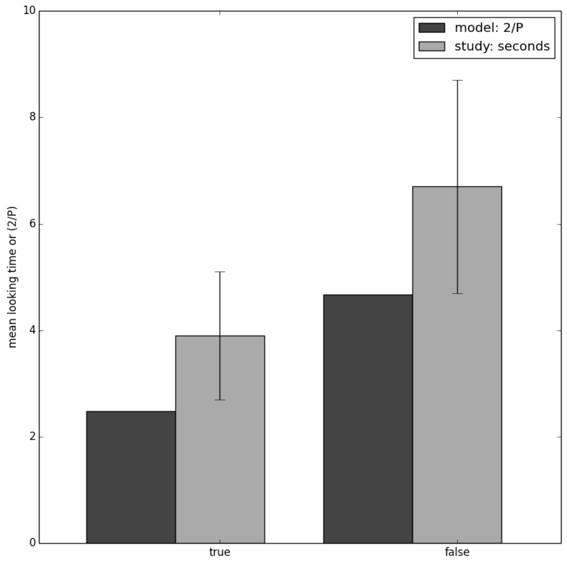


Figure 5.10: Model simulation results for Koenig and Echols (2003). On the Y axis is the mean time in seconds infants spent looking at the informant across trials and for the model, the mean inverse probability of the informants actions across trials.

Apart from the looking-time modifications, the rest of the workings are identical to those we used to model Pasquini et al. (2007). The results can be seen in Figure 5.10. We plot seconds beside inverse probability arbitrarily scaled by 2. The model captures that an informant labeling common objects correctly is less surprising than an informant labeling common objects incorrectly by attributing a high bias toward informants being knowledgeable and helpful (see Table 5.2). Alone, this result suggest that infants expect correct labels. However, Koenig and Echols (2003) ran an additional condition in which an audio speaker sounded object labels. Infants did not look longer at the speaker whether it sounded true or false labels. Together these results suggest that infants specifically expect people to provide correct labels.

5.2.9. Gaze following. Tummeltshammer et al. (2014, Experiment 1) investigated 8-month-olds' learning from informants using a gaze-following paradigm. The researchers employed eye-tracking technology to record infants' eye movements in response to gazes made by *reliable* and *unreliable* faces. For each face type, infants participated in four blocks of four familiarization trials. In each trial, a woman's head appeared in the center of a black screen. In each of the four corners of the screen were empty boxes (squares). At the beginning of each trial the head looked at the infant, said "Wow, look!" and turned to look at one of the four corners at which time an animal noise sounded and its respective animal appeared in one of the boxes. Reliable faces always preemptively looked at the box in which the animal appeared and unreliable faces preemptively looked at the box in which the animal appeared only 25% of the time. Each square had a distinct animal and the heads only ever looked at two of the four boxes. There were two boxes both in which an animal never appeared and which were never looked at. After familiarization trials, infants participated in two different kinds of target trials: *test* and *generalization*. On *test* trials, the head looked at a box it had previously looked at. After a short delay an animal sound played but no animal appeared; instead the corner boxes flashed. The same procedure repeated for *generalization* trails but the head looked at one of the boxes it had never looked at before—the hypothesis in both cases being that if such young infants are sensitive to informant reliability infants who observed the reliable head should be more likely to follow its gaze. In both target trial types, infants looked at the box indicated by the reliable informant far more than the others boxes; infants looked at the box indicated by the unreliable informant at chance.

From a modeling standpoint this study was difficult to capture, not because there is something about it that is inherently difficult to capture, but because the information supplied in the publication does not provide sufficient information to account for all the relevant details.[*] Before the experiment began, infants participated in a number of calibration trials during which objects appear in the corners and center of the the screen. It is possible that these trials affected infants' beliefs about where objects should appear on the screen and hence their learning during familiarization. As an illustration: assume that during calibration infants cumulatively observe ten objects appear in each of the four corners. We capture the likelihood of an object appearing in a given corner with multinomial distribution with Jeffery's prior,

$$P(\text{corner}) \sim \text{Dirichlet}\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right), \tag{5.9}$$

[*] We requested, but were not able to obtain data from the authors

Which is the probabilistic way of establishing a loose, uniform belief that objects are equally likely to appear in any of the four squares. After calibration and posterior probability updates we have

$$P(\text{corner}) \sim \text{Dirichlet}\left(\frac{1}{2}+10, \frac{1}{2}+10, \frac{1}{2}+10, \frac{1}{2}+10\right), \tag{5.10}$$

which amounts to a very rigid uniform belief and which slows future updating—that is to say that each subsequent observation less affects the predictive probability of a specific event. Assuming that infants update their beliefs about objects and corners on each trial, an infant who receives the above calibration trials will attribute a predictive probability of 0.362 to an object appearing in one of the two never-before-indicated boxes on generalization trials where an infant with no calibration trials would attribute a probability of only 0.056 to the same event. We ignored this sort of posterior updating because the study provides insufficient data and, as we have demonstrated, subtle differences in calibration assumptions can lead to dramatically different results. We assume that infants held a uniform probability over objects to corners for the duration of the experiment. It should be noted that there was a qualitative difference in infants' behavior in the two target trials that could be explained by updating beliefs about objects and corners. It appears that infants followed the reliable head's gaze to the cued box more in generalization trials than they did in test trials and followed the unreliable head's gaze less in generalization trials than they did in test trials (see Figure 5.11). If infants are looking for the box with the animal and an unreliable informant looks toward a box in which an animal has never appeared, children should look less because at baseline it is unlikely for an animal to appear there. A reliable head's gaze, to an extent, overrides the low prior probability of an animal appearing in that corner.

Another issue is trial ordering. Just as beliefs about corners and objects propagate across trials, so too do beliefs about informants. The study was conducted using a between-subjects design. The order of the boxes in which the animals appears and—we assume—the order of the trials during which the unreliable informant looked at the correct object were counterbalanced. It is computationally intractable to average over many orderings for an experiment of so many trials, and because we do not have the exact trial orders of each

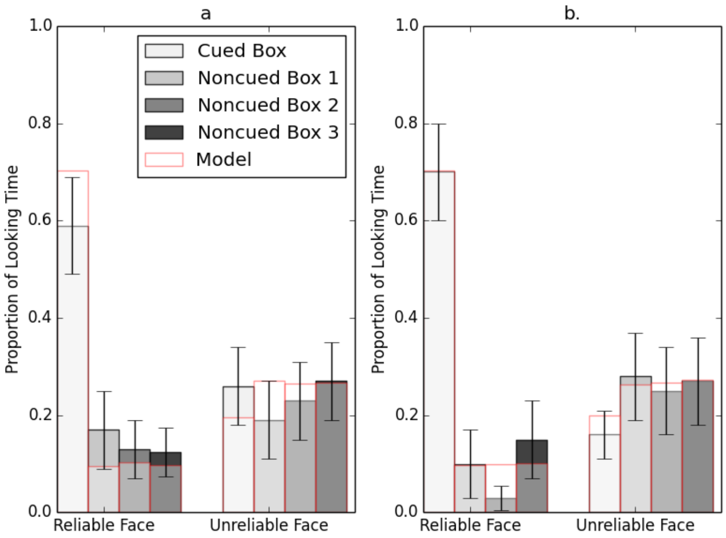


Figure 5.11: Model simulation results for Tummeltshammer, Wu, Sobel, and Kirkham (2014). Error bars represent standard error.

participant, we cannot use approximation methods to capture individuals' behaviors (e.g. win-stay, lose-shift [Bonawitz, Denison, Gopnik, & Griffiths, 2014]). We modeled each condition—reliable and unreliable—separately and chose a single order for the unreliable condition (that the face looked at the correct box on the second trial of each block).

Infants' likelihood of looking at the box indicated by the face was modeled using the same process as modeling an endorse trial. The infant should expect an animal to appear in the box indicated by the face if the face is likely to correctly label (via its gaze) that box as "the box that is going to have the animal in it". In Figure 5.11 we report the model results.*

We see that the model captures infants' preference to follow the reliable face's gaze and to look other than where the unreliable face gazes. Again, there is a qualitative (though not statistically significant) difference in the results for the test and generalization trials for unreliable faces. Infants seems to look uniformly in the test trials (Figure 5.11a) and seem to look other than where the unreliable face looks in generalization trials (Figure 5.11b). Because we have ignored posterior updating with regards to object locations, these two target trials are indistinguishable to the model. Because the two trials are identical to

*Tummeltshammer et al. (2014) did not report their means and did not provide them on request so we used the *ruler-to-bar-chart method* to approximately measure them.

the model, the fits that best explain the aggregated data have a slight, weak bias toward deception (see Table 5.2). This explains the model's preference to not follow the the unreliable face's gaze. The model begins with a bias toward deception and thus is quicker to explain unreliable behavior as deception. Because a deceptive informant is least likely to gaze at the correct box, the model assumes that any of the other boxes are a more likely place for an animal to appear. If the model began with a more neutral bias or perhaps with a pedagogical bias (knowledgeable and helpful), then we might expect more uniform attributions of probability across boxes for the unreliable face. An unknowledgeable face looks uniformly at random. If there is a uniform probability of an animal appearing in any of the boxes then the unreliable face's expected accuracy is 25%–the assumption made by the study. If this is the case, the face would be disregarded entirely because its gaze is not informative.

5.3. Capturing the effect of attachment during infancy on trust

Corriveau, Harris, et al. (2009) looked at the link between infants' attachment with their mothers during infancy and their future willingness to accept information from their mothers over information provided by strangers. They hypothesized that infants would prefer information from their mothers or strangers differently depending on their attachment type. Infants' attachment with their mothers was gaged at 15-months using the *Strange Situation* procedure (Ainsworth, Blehar, Waters, & Wall, 1978). In this procedure a child and caregiver enter a room and the child is intermittently left alone with the stranger or left entirely alone. Coders monitor the exploration behavior of infants, infants' reactions to their caregivers' leaving and returning to the room, and infants' responses to being left alone with the stranger. Infants are divided into four attachment types:

1. **Secure** children comfortably explore the room in the presence of their caregivers. They become upset when their caregivers leave the room but are relieved when they return.

2. **Avoidant** children avoid or ignore their caregivers, and show little reaction to their leaving or returning. They explore very little regardless of who is in the room.

3. **Resistant** children are clingy and show distress upon entering the room.

4. **Disorganized** children are those who do not fit neatly into the other categories. Disorganized infants' behaviors do not appear to be linked to the proximity of the caregiver.

Children and their mothers were then recalled at 50 months (4.17 years) and tested using an ask-endorse paradigm over a variety of conditions: one *novel objects* condition and two *hybrid objects* conditions.

In the four novel objects condition trials, a novel object was placed on a table at which sat the child's mother and a stranger. The child was invited to ask either her mother or the stranger for the object label (ask trial), or after having observed both informants provide labels was asked to label the object herself (endorse trial).

In hybrid objects conditions, instead of novel objects, children were asked about images of hybrid objects, e.g., *bear-pig*, *shoe-car*. There were two levels of hybrid: 50-50 and 75-25. For example, a 50-50 bear-pig would be equal parts bear and pig and it would be just as acceptable to label it a pig or a bear. A 75-25 shoe-car would be more shoe than car and one should be more likely to classify it as a shoe than as a car. Children participated in the same type of ask and endorse trials in which the goal was not to provide a novel label but to classify the object in the image as one of its hybrid parts. For example, to classify the bear-pig as a bear or a pig. There were four trials in each condition.

We do not draw a mapping directly from attachment security to epistemic trust but hypothesize that attachment in part represents a modified expectation on mothers' and strangers' helpfulness and knowledgeability. That is, caregivers provide their infants with experience that may foster epistemic distrust. We shall use the model as crystal balls, asking what it believes in the vaguest terms by fitting the model and observing the hyperparameters $(\alpha_k, \beta_k, \alpha_h, \beta_h)$. To do so we shall construct appropriate probability distributions to model the likelihood of particular beliefs and actions (labels) given a novel object or a hybrid object. The details for novel objects are identical to the previous studies we have modeled (see subsection 5.2.2). For the hybrid objects condition we created a probability distribution under which unknowledgeable informants were more likely to guess correctly as the salience of the major hybrid part increased. For example, an unknowledgeable informant should more likely believe that a 75-25 bear-big is a bear than a 50-50 bear pig and should be more likely to believe that a 25-75 bear-pig is a pig. We constructed the distribution $p(b|w, \neg k)$ from a series of beta distributions with more mass toward the more salient hybrid part and uniform weight for 50-50 hybrid (see Figure 5.12). A linear weight was placed on the mass such that each cell of the distribution was:

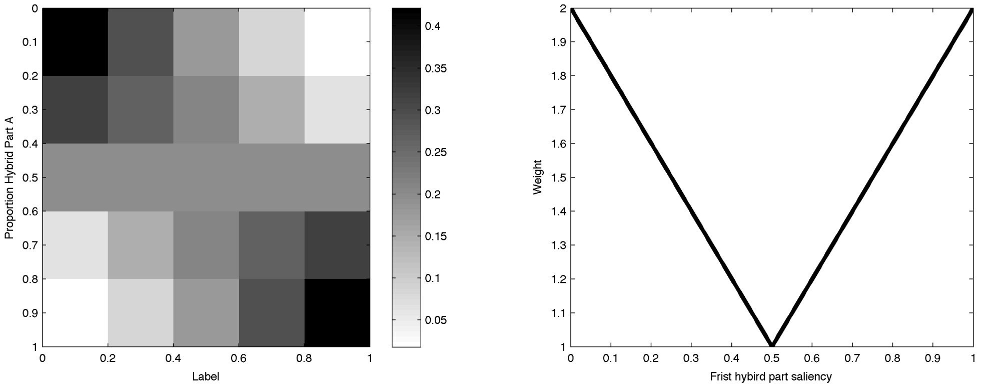
$$P(b|\neg k, w) \propto \begin{cases} \text{beta}\,(1, 2w, 1) & \text{if } w > .5 \\ \text{beta}\,(2(1-w), 1) & \text{if } w \leq .5 \end{cases}, \tag{5.11}$$

where w is the saliency of the first hybrid part and which also corresponds to the true state of the world. This particular scheme is arbitrary and was chosen simply to capture that hybrids with less similar parts should be easier to classify. Here, both the world and belief correspond to the saliency of the first hybrid part. For example a 75-25 hybrid bird-fish implies $b = .75$. Action probabilities, $P(a|b, h)$ were derived in the typical way via the pedagogical sampling recursion (Equation 3.4). We then searched for the best fitting parameters[*] over the grid [0.025, 0.063, 0.16, 0.40, 1., 2.52, 6.33, 15.91, 40.0] (a log-spaced grid from 0.025 to 40). The results can be seen in Figure 5.13.

The correlation between the simulation and empirical results is high, $r(10) = 0.926$, $p < 0.001$. The model captures the different baseline levels of infants' trust in their mothers' novel object labels during novel object trials. The model shows a slight decrease in the strength of the bias during 50-50 hybrid trials. Because both informants provide sensible answers in the 50-50 case—remember that either hybrid part should, by design, be an acceptable label—infants similarly update their beliefs for each informant. However, because infants have more experience with their mothers, new experience more affects their beliefs about the stranger, reducing the bias toward the mother.

Let us work through a simplified example. Assume that the probability that an informant labels correctly has a beta prior,

[*] For this study we searched for the parameters that maximized the likelihood of the results under a binomial distribution.



(a) (b)

Figure 5.12: a) Distribution of unknowledgeable informant beliefs given world states for (hybrid object saliency) hybrid objects.The y-axis represent the salience of the first hybrid part. E.g. .75 for given an image of a 75-25 bird-fish. The x-axis corresponds to the hybrid part of the label. b) Weight function used to derive beta belief distribution. The x-axis is the saliency of the first hybrid part, the y-axis is the weight value. At 50-50 saliency, the weight is 1 causing a uniform distribution of beliefs, i.e., $p(b|w = .5, \neg k) \sim \text{beta}(1, 1)$.

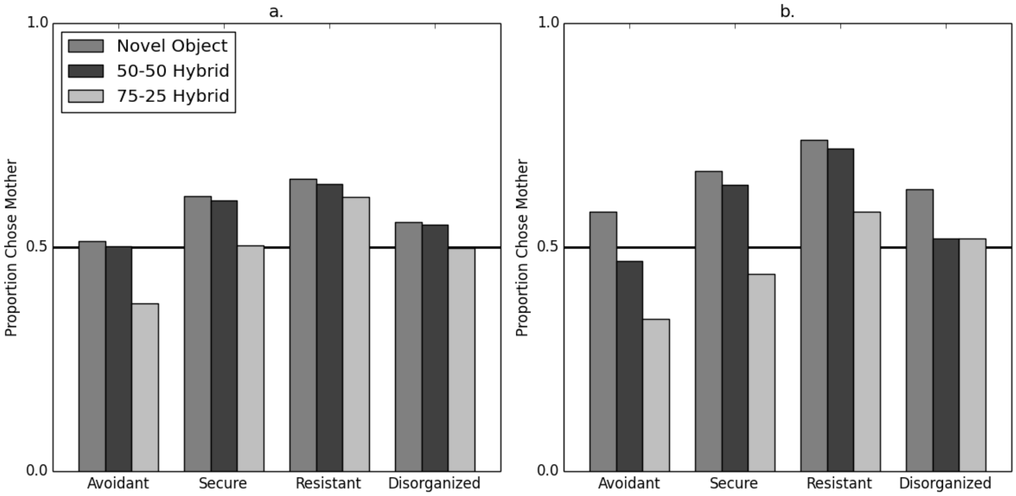


Figure 5.13: Simulated and empirical data for Corriveau, Harris, et al. (2009). a) Model predictions. b) Experimental data. On the y-axis is the proportion of children who endorse the label provided by their mother.

$$p(l = w) = \theta_l, \tag{5.12}$$

$$\theta_l \sim beta(\alpha, \beta). \tag{5.13}$$

Assume that the baseline prior, before any demonstrations, is $\theta_l \sim beta(.5, .5)$. If, out of 20 labellings, the child has observed his mother mislabel only once, the child's expectation that his mother will label correctly in the future is,

$$E[p(l = w)] = E[\theta_l] = E[beta(.5 + 19, .5 + 1)] = \frac{.5 + 19}{.5 + 19 + .5 + 1} = .929. \tag{5.14}$$

While the expected probability with which a novel informant will label correctly at baseline is $\frac{.5}{.5+.5} = .5$. However, during the 50-50 hybrid trials both informants label correctly 4 of 4 times, thus the probability with which the mother will label correctly is,

$$E[p(l = w)] = E[beta(.5 + 19 + 4, .5 + 1)] = \frac{.5 + 19 + 4}{.5 + 19 + 4 + .5 + 1} = .94. \tag{5.15}$$

and the probability with which the novel informant will label correctly is

$$E[p(l = w)] = E[beta(.5 + 4, .5)] = \frac{.5 + 4}{.5 + 4 + .5} = .9. \tag{5.16}$$

so the preference for the mother in our simplified case would have been .65 in novel trials, but .51 after the final 50-50 hybrid trial. In the 75-25 trials the effect is magnified because the mother provides incorrect labels while the stranger provides correct labels. We see this manifest in the results. The preference decreases from novel to 50-50 hybrid and from 50-50 hybrid to 75-25 hybrid.

Table 5.3 shows the best-fitting parameter set derived from the grid search. We see that the parameters for knowledgeability for both mothers and strangers show similar biases for all attachment groups. The model shows priors biased toward knowledgeably ($\alpha_k > \beta_k$)

	Mother				Stranger			
Attachment type	α_k	β_k	α_h	β_h	α_k	β_k	α_h	β_h
avoidant	40.0	0.16	0.025	0.025	6.33	0.063	0.025	0.025
secure	15.91	0.4	0.16	0.025	40.0	0.025	0.025	0.16
resistant	15.91	0.025	0.4	0.025	15.91	0.025	0.025	6.33
disorganized	1.0	0.16	40.0	1.0	6.33	0.063	0.025	0.025

Table 5.3: Best-fitting parameter sets for both strangers and mothers for each attachment level in Corriveau, Harris, et al. (2009).

though the bias is weaker for disorganized-attached beliefs about mothers' knowledgeability. The results appear to be mainly driven by differences in the helpfulness parameters. Most groups show a bias toward mothers being helpful, but avoidant shows uncertainty ($\alpha_h = \beta_h$). Avoidant shows similar uncertainty about strangers' helpfulness. Resistant—the group characterized by clinging to caregivers—shows a strong bias toward deception. That is, resistant results are captured by assuming mothers are helpful and knowledgeable and that strangers are knowledgeable and unhelpful. To reiterate: Avoidant strongly believes that Mother and strangers are knowledgeable but is uncertain about their helpfulness; Secure strongly believes that Mother and strangers are knowledgeable, but weakly believes that Mother is helpful and strangers are unhelpful; Resistant strongly believes that mother and strangers are knowledgeable, but believes that Mother is helpful and strongly believes that strangers are unhelpful; and Disorganized believes that Mother is more knowledgeable than helpful and that strangers are knowledgeable, but is uncertain about strangers' helpfulness. The parameters capture the intuitive notions behind the attachment levels. Avoidant children are uncertain about mothers and strangers, secure children prefer their mothers, resistant children are certain that strangers are unhelpful (malevolent), and disorganized children are none of the above.

5.4. Discussion

In this chapter we demonstrated the capabilities of our epistemic trust model. We have reproduced results from some of our earlier work (Eaves & Shafto, 2012; Shafto et al., 2012) and have modeled new results. We have shown that the model accounts for effects for accuracy, relative accuracy, error magnitude, consensus, pointing, marking, and verbal testimony over a broad developmental range. We have captured not only the results of forced-choice paradigms but also looking-time and eye-tracking paradigms.

We have striven to represent a broad range of literature. In doing so we have had to make a compromise between replicating the data and reducing our degrees of freedom through grouped parameter fitting. We have shown that in general, this has not significantly reduced our ability to capture the data. Importantly we have shown that the model completely fails to account for non-epistemic effects of culture. Recall that the model failed to capture Asian-American children's deference to a majority in the non-affordant tool task (DiYanni et al., under review). The model accounts for results for which it should account and has difficulty accounting for results for which it should not account.

We have also used the model to produce an intuitive account of the effect attachment during infancy has on children's future information-seeking behavior. We showed that given only the task parameters, the prior beliefs that cause the model to best replicate the data are intuitively analogous to the attachment categories.

In chapter 4 we demonstrated that the the model's usefulness extends beyond building rational accounts of empirical results. We used the model's ability to account for an array of results to conduct a computational analysis finding that younger children's results are best explained in the model by attributing a high, strong bias toward helpfulness. This confirmed our previous findings (Shafto et al., 2012) in a natural way, requiring no post-hoc model modifications.

6. AN ARGUMENT AGAINST THE NECESSITY OF AN INNATE PEDAGOGICAL ASSUMPTION

According to Csibra and Gergely (2006), “If our hypothesis about the fundamental role that pedagogy has played during human evolution and plays during human development is correct, this would then also imply that seeing each other as cooperative and omniscient individuals is also part of our nature [. . .] one aspect of social cognitive development will necessarily be to learn when to overcome (suspend or inhibit) these default assumptions[. . .]” (p. 13).

Natural pedagogy paints a very clear picture for development: that people are born with the assumption that informants are knowledgeable and helpful and that a key task in children’s early development is to relax their pedagogical assumption (lest they become vulnerable to misinformation from unreliable informants). In chapter 4 we demonstrated a developmental trend partially in-line with that posed by Natural Pedagogy. The model suggests that development is driven by children’s relaxation of their biases toward believing that all informants are helpful. However, the crux of Natural Pedagogy is that it is innate. To empirically evaluate claims of innateness we must evaluate newborns. Researchers are simply unable to evaluate the complex claims made by Natural Pedagogy on sufficiently young infants to confidently speak to claims of innateness. The looking-time paradigm commonly used by researchers requires very simplified stimulus because it is difficult to control what infants attend to and pedagogy research often relies on elaborate demonstrations with informants and objects. We are reminded of the difficulty in evaluating these complex claims on infants by the large number of participants excluded from analyses in pedagogy experiments.[*] These large exclusion rates are a reflection of the strain placed on

[*] For example, when studying goal attribution in 6.5-month-olds, Csibra (2008) excluded 36% of the initial population; Csibra and Volein (2008) excluded 38% of 8- and 12-month-olds from analyses in their study on inferring the presence of objects from gaze; and Gliga and Csibra (2009) excluded 31% of 12-month-olds

looking-time methods and the delicacy of the specific implementations thereof.

One obvious—albeit logistically and ethically intractable—way to demonstrate Natural Pedagogy is to evaluate a number of feral children raised by non-humans. If humans are innately able to use ostensive cues to engage the pedagogical sampling assumption, the results of Bonawitz et al. (2011) should be reproduced with feral children (of course, if they were not, one could argue that pedagogy is subject to some critical period after which it is irreparably diminished if not used). Single feral children are a rarity, large groups of them doubly so. If we cannot evaluate newborns and lack the subject pool to evaluate feral children, how then can we evaluate the innateness claim of Natural Pedagogy?

In this chapter, we shall use computational simulations to evaluate (and subsequently refute) some claims of Natural Pedagogy by evaluating the benefit of the pedagogical assumption on fledgling learners. First, Natural Pedagogy states that pedagogy is an "adaptation[] to receive knowledge from social partners"(Csibra & Gergely, 2006, p. 1). That is, it improves learning from informants. Second, Natural pedagogy claims that children must learn when to *overcome* the default assumption that people are *omniscient* (knowledgeable) and *cooperative* (helpful). This implies that there is some early benefit to holding this rigid assumptions over other more robust assumptions(Csibra & Gergely, 2006). We shall demonstrate that:

1. Natural Pedagogy does not necessarily offer optimal learning compared with other weakly-biased models, regardless of informant type.

2. If all informants are teachers alternative learning models learn about the world at similar rates.

Thus demonstrating that the pedagogical sampling assumption, to naive learners, is at best a burden without benefit and is otherwise suboptimal.

in their study on, what they refer to as, *deictic gestures* (pointing or looking as indications an object is present).

6.1. Methods

We conduct a set of temporal simulations consisting of interactions between informants and a naive learner (a learner who has no knowledge). Across interactions, we monitor learners ability to learn about informants and about the world. We compare the learning performance of learners under the Natural Pedagogy assumption to that of learners endowed with various learning models alternative to Natural Pedagogy.

Each interaction of the simulation proceeds roughly as follows:

1. A single informant provides a demonstration in which she labels an object or elicits an effect from a causal system through intervention.
2. The learner uses the evidence (the observed action and effect) and her knowledge of the world and the informant (is she has any) to infer the distribution of the informant's helpfulness and knowledgeability and to learn the world (if she does not already know it).
3. The learner uses what she has learned about this informant to update her beliefs about informants in general.

6.1.1. Alternative models. The Natural Pedagogy assumption corresponds to a learner who can represent both informants' helpfulness and knowledgeability but has strong biased beliefs that all informants are helpful and knowledgeable. We derive several alternatives to Natural Pedagogy by relaxing the biases and pruning nodes from the full epistemic trust model (see Figure 6.1). A helpfulness-only model makes inference based only on helpfulness, a knowledgeability-only model makes inferences based only on knowledgeability, and a naive model make assumptions based only on the data and effects. To create a helpfulness-only model, we remove the knowledgeability node and its parents, remove the beliefs node, and connect the world node to the action node (there is no need to represent beliefs if knowledgeability is not represented); to create a knowledgeability-only model, we remove the helpfulness node and its parents; to create a naive model, we remove both the helpfulness and knowledgeability nodes and their parents. We simulate learning under strong biases toward knowledgeability and helpfulness,

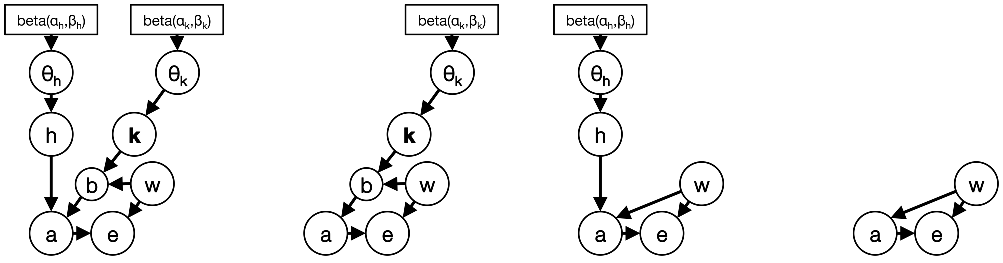
$$\theta_k \sim \text{beta}(10.5, .5) \tag{6.1}$$

$$\theta_h \sim \text{beta}(10.5, .5), \tag{6.2}$$

and neutral, noninformative biases (based on the commonly-used Jeffrey's Prior [Jeffreys, 1946]),

$$\theta_k \sim \text{beta}(.5, .5) \tag{6.3}$$

$$\theta_h \sim \text{beta}(.5, .5). \tag{6.4}$$



(a) Full model. (b) k-only model. (c) h-only model. (d) Naive model.

Figure 6.1: Models used for simulations. a) Full model as described in chapter 3. b) Knowledgeability-only (k-only) model in which all informants are implicitly helpful. c) Helpfulness-only (h-only) model in which all informants are implicitly knowledgeable. Note that the belief node has been removed and the world node has been routed to the action node. d) The naive model in which informants act only to demonstrate the true state of the world.

6.1.2. Simulating informants. Simulating interactions between learners and informants requires simulating informants. Informants draw data according to the full generative model. That is, informants' variable knowledge about the world determines their beliefs; actions are, in turn, determined by informants' beliefs and variable helpfulness. To simulate differently-reliable informants we assign variable probabilities of knowledgeability and helpfulness (θ_k and θ_h) to informants. For these simulations we use four types of informants (see

Table 6.1): a teacher who is most often helpful and knowledgeable, an ignorant informant who is most often unknowledgeable and helpful, a deceptive informant who is most often knowledgeable and unhelpful, and an erratic informant who is most often either knowledgeable or helpful but less often both. These informants occur with variable probabilities depending on the specific simulation. The default (aribitrarily–chosen) assumption is that 70% of informants are teachers and the other 30% are another type with uniform probability. This captures the notion that most informants are helpful and knowledgeable and others informants have a variety of different epistemic qualities. We shall also investigate learning when all informants are teachers.

type	θ_k	θ_h	$P(k, h \vert \theta_k, \theta_h)$	$P(\neg k, h \vert \theta_k, \theta_h)$	$P(k, \neg h \vert \theta_k, \theta_h)$	$P(\neg k, \neg h \vert \theta_k, \theta_h)$
teacher	.95	.95	.903	.048	.048	.003
ignorant	.15	.95	.143	.808	.008	.043
deceptive	.95	.15	.143	.008	.808	.043
erratic	.6	.6	.360	.240	.240	.160

Table 6.1: Informant types used for simulations and their probabilities of helpfulness and knowledgeability. $P(k, h|\theta_k, \theta_h)$ indicates the probability with which the informants is helpful and knowledgeable, $P(\neg k, h|\theta_k, \theta_h)$ indicates the probability with which the informant is not knowledgeable but helpful, $P(k, \neg h|\theta_k, \theta_h)$ indicates the probability with which the informant is knowledgeable but not helpful, and $P(\neg k, \neg h|\theta_k, \theta_h)$ indicates the probability with which the informant is neither knowledgeable nor helpful.

6.1.3. Learning situations. Informants will provide two types of learning situations: labeling and causal. These two situations have different implications for learners. In causal situations, informants intervene on a causal system and learners make inferences about informants based on their ability to learn about the system given the intervention and the elicited effect. In labeling demonstrations, informants provide labels for objects with which learners may or may not be familiar. Causal learning requires less trust in informants than does label learning. Causal demonstrations, unlike labeling demonstrations, result in effects which are veridical data directly from the world. Additionally, object labels, unlike cause–effect relations, are purely conventional. In this respect, learners rely on others to produce labels in a conventionally appropriate way.

We used causal structures of only two or three nodes. To create causal structures, we generated all directed acyclic graphs (DAGs) on two and three nodes.[*] Informants act on the structure by activating a node. The activation of that node follows the edges and activates each connected node according to a Noisy-OR relationship in which a child node is activated if any of its parent nodes are activated. For inference purposes, learners know how many nodes are in a structure, they need only to learn the structure's configuration (which nodes are connected to which other nodes).

6.1.4. Simulating learners. Learning under the model is a joint inference problem: learn about the world while learning about informants. Learners use evidence (the observed action and effect) and their knowledge of the world to update their prior beliefs about informants. Learners use their beliefs about informants to make more enlightened inferences about the world given informants' actions. In this way, learning about the world improves one's ability to learn about informants and learning about informants improves one's ability to learn about the world. Learners *learn to learn.* Updating a set of parameters (beliefs) given data (input from informants and the world) in our Bayesian network framework is a simple matter and works in the same way: alternating between updating prior beliefs on informants' knowledgeability and helpfulness given data and learning about the world given the updated beliefs. As in chapter 3 the learner uses the data provided by the informant and the world (the action, effect, and the true state of the world if the learner knows it) to infer the distribution of that informant's probability of being helpful or knowledgeable, $p(\theta_k, \theta_h | a, w)$. Because we cannot calculate this function analytically we use sampling methods to generate many values of θ_k and θ_h and then estimate the beta distribution from which they were generated via the method of moments.[†] If a new informant is encountered the default values of α_k, β_k, α_h, and β_h are generated by averaging the estimated distributions for each existing informant by combining draws from each informants' distribution and estimating a new distribution from those draws using the method of moments.

Learners begin with no knowledge of the world. They cannot leverage their knowledge of

[*] We chose 2 and 3 nodes because the number of labeled DAGs grows prohibitively quickly. The number of DAGs on 2 nodes is 3, the number DAGs on 3 nodes is 25, the number of DAGs on 4 nodes is 543; the number of DAGs on 5 nodes is 29281 (R. W. Robinson, 1977).

[†] Given a set, X, of draws from a beta distribution with sample mean, $\bar{x}$, and sample variance $\bar{v}$, the estimated beta distribution is beta$(\hat{\alpha}, \hat{\beta})$ where $\hat{\alpha} = \bar{x}S$ and $\hat{\beta} = (1 - \bar{x})S$ and $S = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right)$.

the world to learn about informants and thus must rely on the data produced by informants to learn about the world. If at the end of an interaction the learner attributes the highest probability to the correct state of the world, $\text{argmax}_w(P(w)) = w_{true}$, the learner is said to have learned that state of the world and may use that information in future interactions.

6.1.5. Procedure. The simulation occurs in two parts. The world, and informants and their actions are generated (pseudocode in Algorithm 1). These data are then sequentially fed to learners who then learn about the world and informants (pseudocode in Algorithm 2).

Algorithm 1 Initialize learning environment

```
Inputs: Set of world states, $W$, and number of informants $I$ or expected number of informants, $E[I]$
Outputs: Set of world states, worlds; set of informants for each trial, informants; actions and effects for each trial, evidence
worlds ← empty list
informants ← empty list
evidence ← empty list
for t in $0, 1, \ldots, T-1$ do
    worlds[$t$] ← choose random world state, $w \in W$
    if constant number of informants, $I$ then
        informants[$t$] ← choose random informant, $i \in I$
    else_expected number of informants, $E[I]$
        $p$ ← random float $\in [0, 1)$
        if $p < E[I]/T$ then
            informants[$t$] ← new informant
        else
            informants[$t$] ← choose random informant, $i \in I$
        end if
    end if
    knowledgeable ← True with $P(\text{knowledgeable}) = \text{informants}[t].\theta_k$
    helpful ← True with $P(\text{helpful}) = \text{informants}[t].\theta_h$
    belief ← draw from $\text{belief}|\text{worlds}[t], \text{knowledgeable}$
    action ← draw from $\text{action}|\text{belief}, \text{helpful}$
    effect ← draw from $\text{effect}|\text{worlds}[t], \text{action}$
    evidence[$t$] ← $\{\text{action}, \text{effect}\}$
end for
return informants, worlds, evidence
```

Algorithm 2 Learn from evidence

```
**Inputs**: Output of learning environment initialization, informants, worlds, evidence; starting prior parameter sets for $k$ and $h$, defaultParams$_k$ and defaultParams$_h$.
**Outputs**: Number of worlds learned at trial $t$, numWorldsKnownTrial; default parameter set at trial $t$, paramsAtTrial numWorldsKnown ← 0 numWorldsKnownTrial ← empty list paramsAtTrial ← empty list
for $w \in W$ do
    knowsWorld[$w$] ← False                                ▷ Learners begin with no knowledge.
end for
if constant number of informants, $I$ then
    for $i \in I$ do
        informants$_l$[$i$].params$_k$ ← defaultParams$_k$                  ▷ set $\{\alpha_k, \beta_k\}$
        informants$_l$[$i$].params$_h$ ← defaultParams$_h$                  ▷ set $\{\alpha_h, \beta_h\}$
    end for
else_expected number of informants, $E[I]$
    informants$_l$[0].params$_k$ ← defaultParams$_k$
    informants$_l$[0].params$_h$ ← defaultParams$_h$
end if
for t in $0, 1, \ldots, T-1$ do
    i ← informants[$t$]
    if i $\notin$ informants$_l$ then
        informants$_l$[$i$].params$_k$ ← defaultParams$_k$
        informants$_l$[$i$].params$_h$ ← defaultParams$_h$
    end if
    params ← $\{$informants$_l[i]$.params$_k$, informants$_l[i]$.params$_h\}$
    $w_{true}$ ← world[$t$]
    if knowsWorld[$w$] then
        $\theta_k, \theta_h$ ← draw from $\theta_k, \theta_h|$evidence$[t]$, params, $w_{true}$
    else
        $\theta_k, \theta_h, p_w$ ← draw from $\theta_k, \theta_h|$evidence$[t]$, params
    end if
    informants$_l$[$i$].params$_k$ ← methodOfMomentsEstimate($\theta_k$)
    informants$_l$[$i$].params$_h$ ← methodOfMomentsEstimate($\theta_h$)
    defaultParams$_k$, defaultParams$_h$ ← averageInformantParamters(informants$_l$)
    if  not knowsWorld[$w$] and argmax$_w(p_w) = w_{true}$ then
        knowsWorld[$w$] ← True
        numWorldsKnown++
    end if
    numWorldsKnownTrial[$t$] ← numWorldsKnown
    paramsAtTrial[$t$] ← $\{$defaultParams$_k$, defaultParams$_h\}$
end for
return numWorldsKnownTrial, paramsAtTrial
```

6.2. Results

Each simulation was run for 500 iterations and had 50 possible world states and either a fixed number of 50 informants or informants were introduced probabilistically such that the expected number if informants was 50.* A summary of the results can be found in Table 6.2.

I	$E[I]$	Prior bias	Informant types	Qualitative results
50	·	neutral	normal	(Full, h-only),(k-only, naive)
·	50	neutral	normal	(Full),(h-only),(k-only, naive)
·	50	strong	normal	No difference between models
·	50	strong	teachers	No difference between models
·	50	neutral	teachers	No difference between models

Table 6.2: Simulation results reference. I indicates the number of informants if the number of informants was fixed. $E[I]$ indicates the expected number of informants if the number of informants was not fixed. Prior bias indicates whether the learner began with a *strong* bias toward believing all informants are helpful and knowledgeable or a *neutral* or non-informative bias. Informant types indicates the types of informants to which the learner was exposed: *teachers* indicates that all informants were teacher; *normal* indicates that 70% of informants were teachers and 30% were chosen randomly among the other types. The rightmost column is a summary of the world–learning results. Listed from best–performing to worst–performing. Similarly performing models are grouped by parentheses.

6.2.1. Learning. We compared whether introducing informants all at once or one at a time had an effect of learning. For the first set of simulations, 70% of the simulated informants were teachers the rest were uniformly distributed among the other types (see Table 6.1). This corresponds to the idea that most, but not all, informants are helpful and knowledgeable. Figure 6.2 shows learning given a fixed number (50) informants and Figure 6.3 shows learning given that new informants are introduced with probability .1 such that the expected number of informants is 50. We see that the end point with respect to learning the world (Subfigure a) is similar under both assumptions. The full model performs best, the h-only model outperforms the k-only model, which performs similarly to the naive model. The k-only model performs worse than the h-only model because it does not account for noise (informants' deviance from optimal, teaching actions) for multiple types of informants as robustly as the h-only model. The k-only model attributes noise system-

[*] For a simulation of $T = 500$ trials, to create an expected number of $E[I] = 50$ informants, we introduce a new informant on each trial with probability $E[I]/T = 50/500 = 0.1$

atically by attributing false beliefs to the informant. Informants who are deceptive hold correct beliefs but nonetheless produce suboptimal data. The h-only model acts as statical learning with variable noise; it simply learns that certain informants choose sub-optimally according to some probability but does not attribute noise to an epistemic attribute.

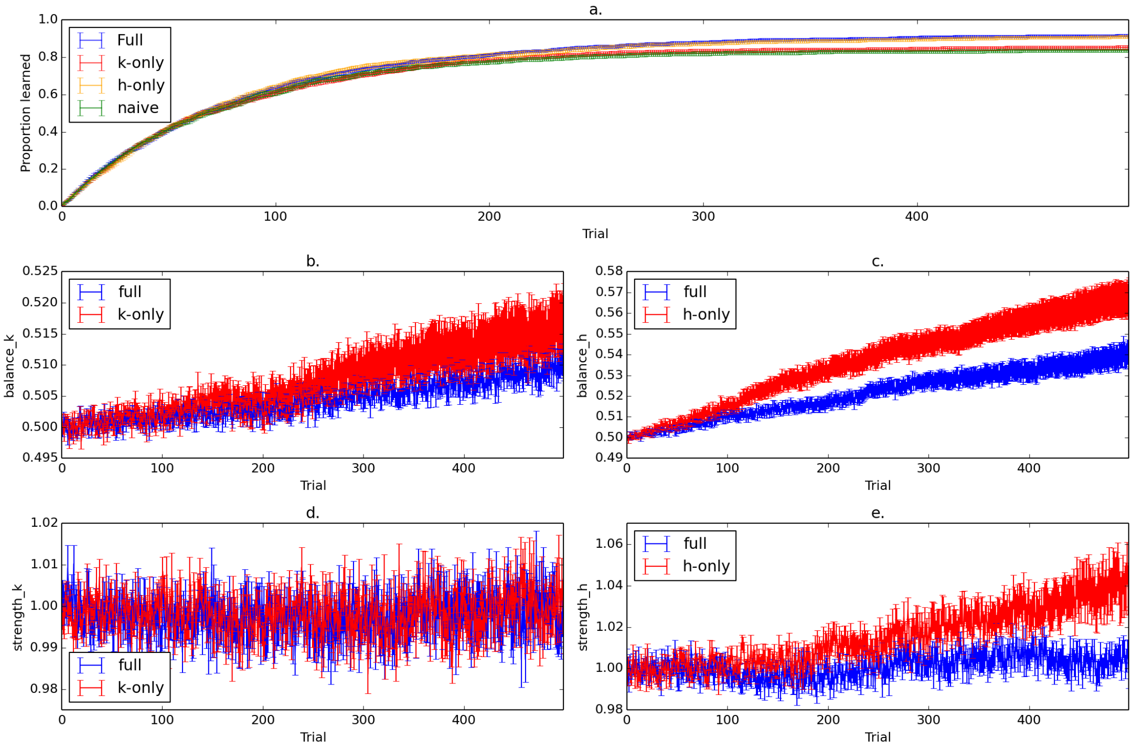


Figure 6.2: Mean learning from 70% teachers given neutral starting beliefs and a fixed number ($I = 50$) of informants over 50 trials. X-axes for all panels displays the trial number. a) Proportion of the world learned at trial for each alternative model. b) Average inferred b_k (balance) parameter for the full (blue) and k-only (red) models. c) Average inferred b_h parameter for the full (blue) and h-only (red) models. d) Average inferred s_k (strength) for the full (blue) and k-only (red) models. e) Average inferred s_h for the full (blue) and h-only (red) models.

The different number–of–informant assumption gives rise to different behavior in learners' learning about informants. Fixing the number of informants (Figure 6.2b-e) leads to more linear learning curves as learners sequentially update their beliefs about existing informants. Learners who are provided new informants as trials progress (Figure 6.3b-e) learn more quickly in early trials as they generalize what they have learned about existing informants to future informants. Models learning given a fixed number of informants behave as expected: they increase their balance parameters toward helpfulness and knowledgeability.

The expected–number–of–informants simulations show different patterns. Most notably, we see in Figure 6.3b that the k-only model decreases its balance for knowledgeability which it then begins to recover toward the end of the simulation. The h-only model's balance for helpfulness also dips but reverses quickly (Figure 6.3c). This suggests that the k-only model has more difficulty generalizing to multiple informant types. We will focus only on simulations in which new informants are introduced probabilistically (the more lifelike assumption) because in general, across simulations the world–learning end–points are similar for fixed and expected number of informants.

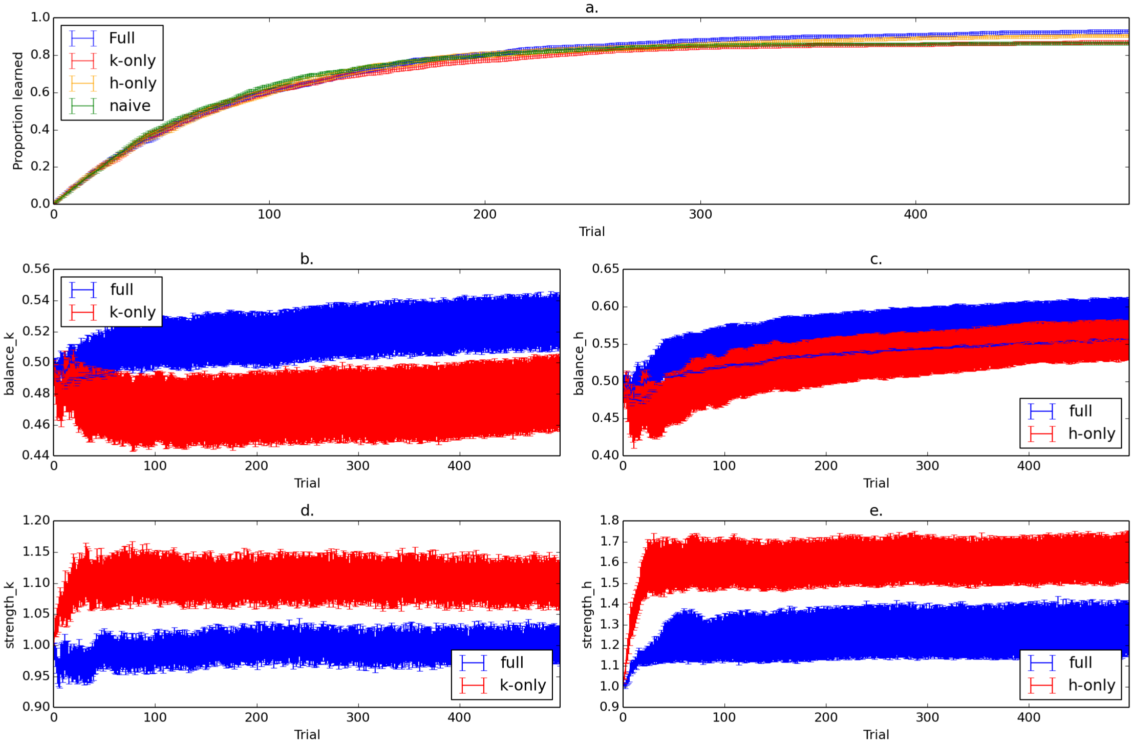


Figure 6.3: Mean learning from 70% teachers given neutral starting beliefs and a variable ($E[I] = 50$) number of informants over 50 trials. X-axes for all panels displays the trial number. a) Proportion of the world learned at trial for each alternative model. b) Average inferred b_k (balance) parameter for the full (blue) and k-only (red) models. c) Average inferred b_h parameter for the full (blue) and h-only (red) models. d) Average inferred s_k (strength) for the full (blue) and k-only (red) models. e) Average inferred s_h for the full (blue) and h-only (red) models.

Figure 6.4 shows learning when learners begin with strongly biased beliefs that informants are knowledgeable and helpful. In this simulation, the full model corresponds to the pedagogical assumption. In short: all models perform similarly. A model that has a strong belief that all informants are knowledgeable and helpful—the pedagogical assumption—performs similarly to the naive model—the statistical learning assumption. The full, h-only, and k-only models will continue to perform similarly to the naive model until they have sufficient data to account for noise in informants. That means that early on, Natural Pedagogy is needlessly complicated. By the 500th trial, all parameterized models (full, h-only, and k-only) remain strongly biased toward believing that informants are generally teachers and the rate of relaxation of these biases seems to slow as trials progress (see Figure 6.4b-e). Importantly, we see that the neutrally–biased full model and h-only model in Figure 6.3 both fare better than the natural pedagogical model (strongly-biased full model) in Figure 6.4a. This suggests that if people were born with a complex internal learning model that accounts for informants' knowledgeability and helpfulness they would be better served by a model with more flexible biases than those proposed by Natural Pedagogy. Figure 6.3 demonstrates that beginning with neutral biases allows learners to more quickly adapt to their informants and leads to better learning.

A counter argument may be that 70% teachers is a rather dour assumption. It may be that all of infants' early informants are teachers. Figure 6.5 shows that if all informants are teachers that each model performs similarly. That is, if all informants are teachers, the full model—regardless of its starting biases—offers no advantage over statistical learning.

6.2.2. Learning under causal and labeling trials. We also looked at whether learning was significantly different for causal and labeling trials. We ran simulations for 500 iteration with an expected number of 50 informant which were 70% teachers and 30% other. Learners began with either biased beliefs ($\theta \sim \text{beta}(10.5, .5)$) or with neutral beliefs ($\theta \sim \text{beta}(.5, .5)$).

Figure 6.6 shows the learning outcomes for each of the four situations. Learning is similar for causal and labeling demonstrations when learners begin with biased beliefs. As mentioned before, having strong beliefs that all informants are knowledgeable and helpful makes it difficult to learn from informants who are not. The trends between causal and labeling trials are different for learners with neutral starting beliefs. For word learning, the full and helpfulness-only models perform similarly; the knowledge-only models performs worse, and the naive model performs the worst. However, for causal trials, the full model

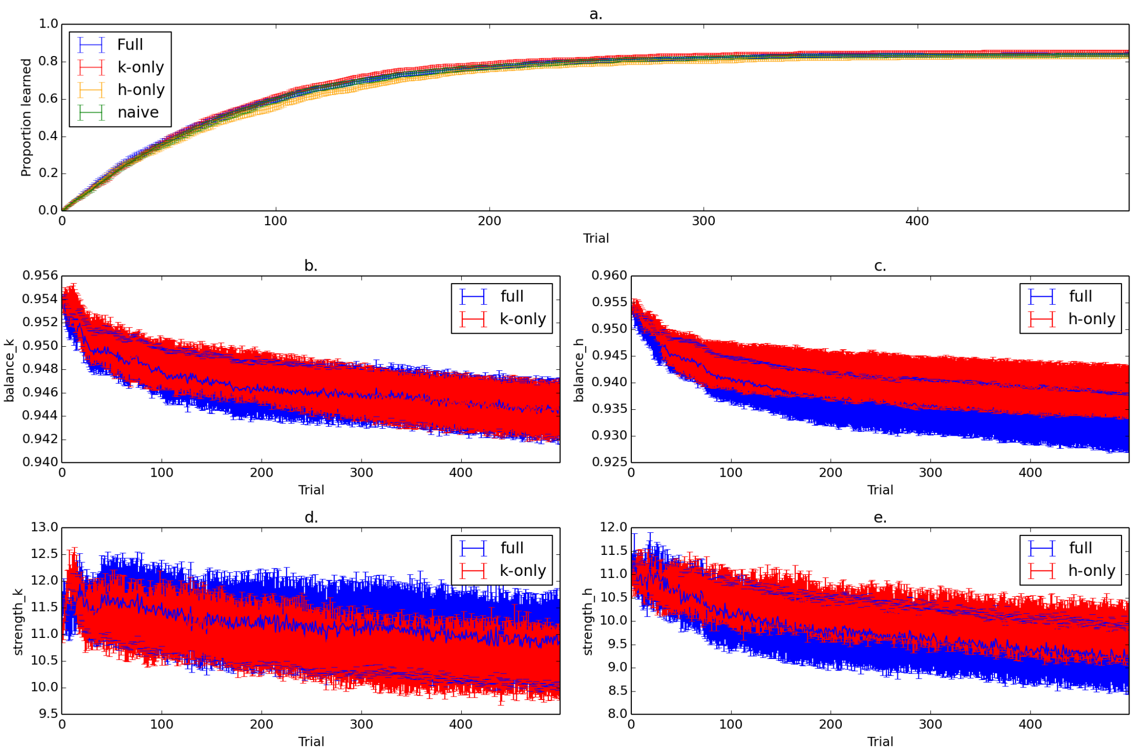
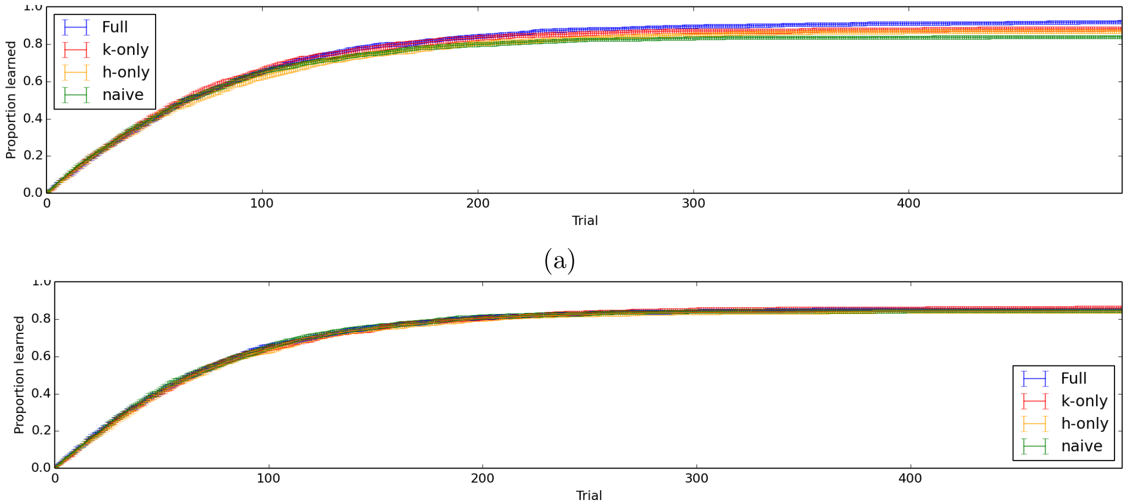


Figure 6.4: Mean learning from 70% teachers given strongly biased starting beliefs and a variable ($E[I] = 50$) number of informants over 50 trials. X-axes for all panels displays the trial number. a) Proportion of the world learned at trial for each alternative model. b) Average inferred b_k (balance) parameter for the full (blue) and k-only (red) models. c) Average inferred b_h parameter for the full (blue) and h-only (red) models. d) Average inferred s_k (strength) for the full (blue) and k-only (red) models. e) Average inferred s_h for the full (blue) and h-only (red) models.

outperforms the others, which show similar performance. It appears that the ability to learn about informants is more important in learning from causal trials because the hypothesis spaces are larger and the effects are subtle. A knowledgeable informant can easily produce an effect that rules out many alternative causal structures or can create effects that lead learners to the wrong structure; unknowledgeable informants interventions must be, to some extent, disregarded.

6.3. Limitations

As is often the case with simulations of this kind, we have had to make a number of simplifying assumptions. We have had to make assumptions about the number and type of informants children interact with. We chose a large number of informants and wide variety



(a)

(b)

Figure 6.5: Learning given only teachers. X-axes show trial number, y-axes show the proportion of hypotheses learned. a) Learning rate for learners with neutral starting beliefs. b) Learning rate for learners with strongly biased starting beliefs.

of informant types to allow us to capture inferences about informants in general (rather than inferences about specific informants) and to be as robust a possible.

We also assumed that a learner has learned something if he assigns the highest probability to that something. This means that a learner has only learned something if it is the truth. This precludes the possibility of a learner maintaining a false belief. If in our simulations we are to allow learners to maintain false beliefs we must provide learners with a way to rectify their false beliefs. We avoided this because there is insufficient data in the literature to model how people override their beliefs given new data. Some research has used minimal groups paradigms (Elashi & Mills, 2014) and Asch tasks (Asch, 1956; Kim, Song, Corriveau, & Harris, 2013; Corriveau & Harris, 2010) to demonstrate situations in which people may override their perception or rationality. However, as we showed in subsection 5.2.5, deferring to one's group or a majority are likely non-epistemic motivations and it is unlikely that learners are actually overriding their beliefs.[*] The key findings are likely to hold regardless of whether learners hold false beliefs because: 1) Strongly biasing the different variations of the model toward believing that all informants are teachers makes them roughly equivalent 2) If all informants are teachers the number of false beliefs learn-

[*] It would be simple to conduct an experiment to test children's adherence to the axiom *today's posterior is tomorrow's prior* with respect to the true state of the world.

ers acquire through a mistake by a teacher (noise) shall be minimal and shall be quickly overridden by subsequent demonstrations.

Natural Pedagogy claims that data provided in pedagogical contexts result in more generalizable inferences. We are not able to comment on this proposition. The model captures the idea of a stronger inference but we do not currently model inferences that span multiple contexts or domains (e.g. that a tool has multiple purposes) that establish a generalizable inference.

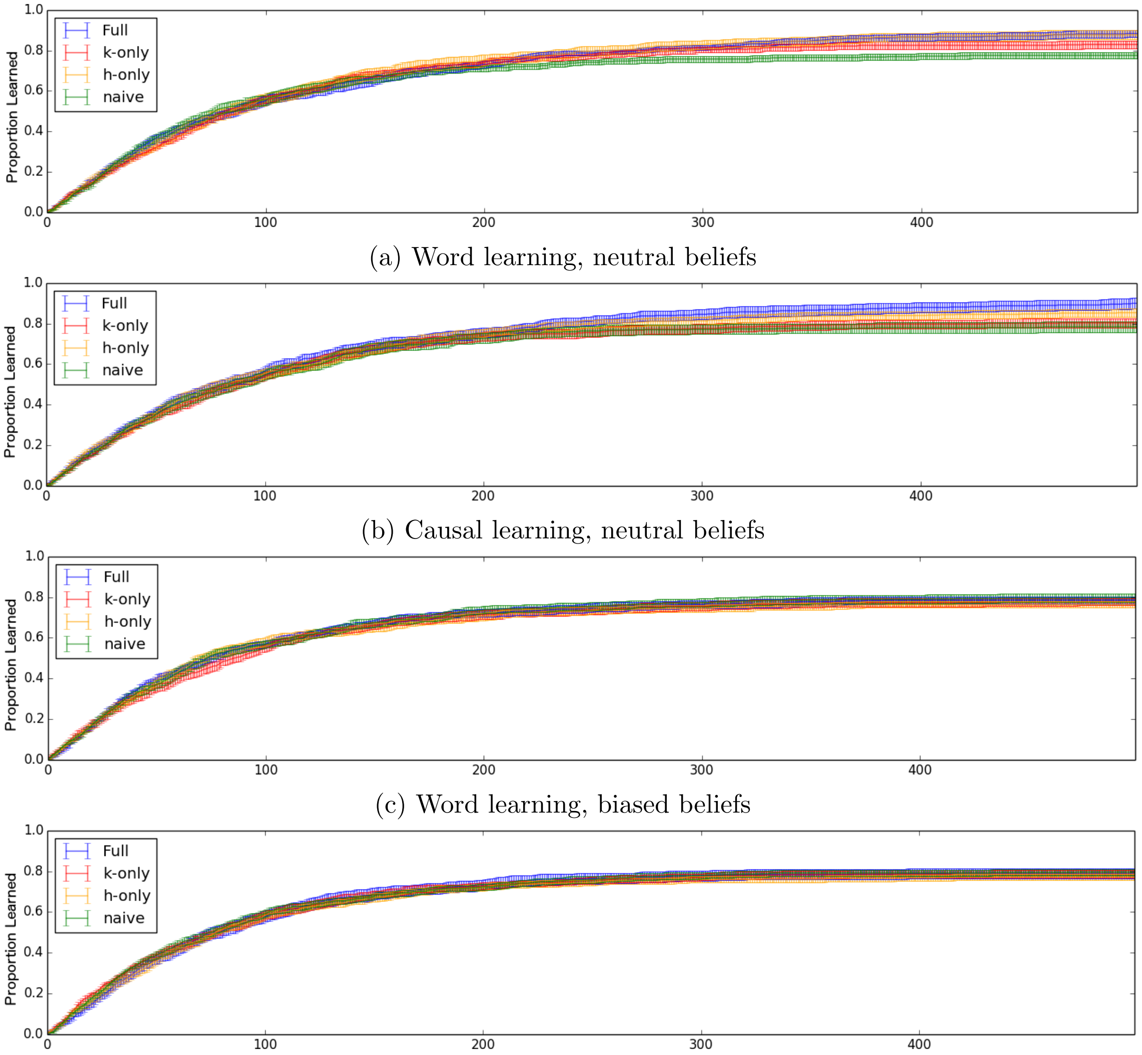
6.4. Conclusion

In this chapter we have constructed simulations that demonstrate that the Natural Pedagogy assumption—that learners assume informants are helpful and knowledgeable—does not offer any early advantage over statistical learning. We have also shown that the complexity of the full model does not afford improved learning over statistical learning if all informants are teachers. Thus, from an efficiency standpoint, it is unlikely that early learners represent informants' helpfulness and knowledgeability but learn that informants have knowledgeability and helpfulness later on. To explain, the pedagogical assumption is immutably built into the naive model. If all informants are teachers there is no reason not to be a statistical learner. It is only when a learner routinely encounters informants who do not behave as teachers, or when the knowledge domains learners confront go beyond common knowledge, that considering informants' epistemic qualities (their helpfulness and knowledgeability) becomes important to learning.

Another point against Natural Pedagogy is that it does not result in development in–line with the literature. We observed in chapter 4 that development is driven mainly by a relaxation of the helpfulness assumption. In Figure 6.4b-e (blue) we plot the development of trust given the pedagogical assumption. We see that the helpfulness bias is unlearned more quickly than the knowledgeability bias. This implies that if the pedagogical assumption is innate, children should very quickly develop the ability to recognize deceptive informants, but this is not what we see in the literature. Younger children have less difficulty making judgments about accuracy when informants' knowledgeability is left variable (compare

Pasquini et al., 2007; Couillard & Woodward, 1999).

A more likely explanation is that infants incrementally add complexity to their data–selection models as the data necessitate. Learners may first learn to attribute inaccuracies to informants' lack of knowledge and may later learn that knowledgeable informants may produce inaccurate data through their variable intentions. The literature indicates that children have the ability to monitor knowledgeability at a very early age (Tummeltshammer et al., 2014; Koenig & Echols, 2003) but our analyses suggest that children's ability to monitor helpfulness begins to develop around the age when we are just able to study them. As it stands now, we simply do not have the data to use the model to make specific statements about development this early. In the future we hope to see more data from one– and two–year–olds which we can use to evaluate these proposals.



(a) Word learning, neutral beliefs

(b) Causal learning, neutral beliefs

(c) Word learning, biased beliefs

(d) Causal learning, biased beliefs

Figure 6.6: Learning rates for causal-only and word-only simulations. X-axes show trial number, y-axes show the proportion of hypotheses learned. Y- a) Learning rate for learners over labeling trials for neutral starting parameters, $\theta_k \sim$ beta(.5,.5), $\theta_h \sim$ beta(.5,.5). b) Learning rate for learners over causal trials for neutral starting parameters. c) Learning rate for learners over labeling trials for biased starting parameters, $\theta_k \sim$ beta(10.5,.5), $\theta_h \sim$ beta(10.5,.5). b) Learning rate for learners over causal trials for biased starting parameters.

7. CONCLUSION

Our goal for this work was to create a continuous account of the development of trust and to use simulation to speak to the innateness claims of Natural Pedagogy. We began by addressing the issue of input sufficiency for infants to learn that people produce data pedagogically (knowledgeably and helpfully) in certain contexts. If infants do indeed learn that people choose data pedagogically they appear to do so early. Research clearly demonstrates that young infants have different expectations for informants (Koenig & Echols, 2003; Topal et al., 2008; Tummeltshammer et al., 2014). To learn that people choose data differently in certain contexts, a statistical learner must be able to identify those contexts and observe differences in data in multiple contexts. We identified infant-directed speech (IDS) as early input from which children may potentially begin to learn social learning. Infants are bombarded by speech. They certainly observe normal conversational, adult speech but adults also provide infants with their own special brand of speech. We have shown through computational modeling, that IDS shares features in common with pedagogical speech—speech optimized to teach phonetic categories. This suggests that infants may receive sufficient data sufficiently early in development that a pedagogical assumption could develop before an age when experimenters could reliably evaluate infants without it.

Any presumption, innate or learned, that people are knowledgeable and helpful must then be unlearned. This phenomenon is the subject of the epistemic trust literature where children's behavior is commonly explained in terms of tracking who is *knowledgeable*. We formalized a probabilistic account of learning from informants which takes into account both informants' variable knowledgeability and their variable helpfulness. We used the model to account for a wide array of research spanning a broad age range and outlined a framework for computational meta-analyses. The meta-analysis indicated that trust development, at least within the ages of the children in the studies we modeled, was driven mainly by relaxation of the helpfulness assumption. We observed that the model captured younger children's data with a higher, stronger bias toward believing that all informants are helpful and knowledgeable. This reproduced our previous findings but in an organic way which did not require us to impose constraints on the model. The success of the model provides evidence connecting the pedagogical learning and epistemic trust literatures, provides support for developmental changes in reasoning about helpfulness (not knowledge), and introduces a powerful new methodology for performing joint analysis of studies that differ on multiple dimensions as is common in development.

Finally, we conducted simulations to evaluate some of the innateness claims of Natural Pedagogy. The Natural Pedagogy theory states that the pedagogical assumption (that all informants are knowledgeable and helpful) is an innate bias which evolved to improve learning. We initialized a learner who knew nothing about the world and simulated learning under different learning models. We found that when informants are not always teachers (helpful and knowledgeable), strong biases that all informants are helpful and knowledgeable lead to poorer learning than neutral biases. We also found that if all informants are teachers, the pedagogical assumption provides no benefit over a naive model akin to statistical learning.

Psychologists hold behavioral experimentation in high regard and rightly so. Precise experimental manipulations allow us to query the world for unbiased evidence and replicate phenomena. Less regard is paid to the complementary set of tools. Precise computational formalizations allow us to query our theories, to find out what data they can and cannot explain, and replicate predictions. In an ideal world, these tools would work together, hand in glove, to allow us to precisely gather and precisely interpret evidence.

This document exposes the limitations imposed on even quite brilliant researchers when practicing imprecise, intuitive theorizing. Researchers in the infant directed speech literature have intuitively argued that any two categories or any increase in variance would be evidence that IDS is suboptimal for learning. The computational formalization shows that this is in fact a predicted feature of input optimized to teach phoneme categories. Similarly, the literature on epistemic trust characterized developmental shifts as due to changes in the ability to monitor knowledge. The computational model shows that these changes are best explained in terms of changes in helpfulness. Finally, researchers have argued that, to explain the remarkable speed of early learning, children must come endowed with an innate assumption that others are knowledgeable and helpful selectors of evidence. Computational simulations again show that this is not so: the evidence suggests that assuming that people are knowledgeable and helpful is at best unnecessary and at worst harmful.

Math is pitiless and unforgiving. Building a mathematical model has a way of magnifying weakness in our theoretical understanding of the phenomena we wish to model. Fortunately, the modeling process often makes these weakness explicit. This is one of the ways that mathematical models instruct future research. Just as precise behavioral experiments are viewed as an invaluable tool in understanding human learning, it is my hope that this work would lead at least some to see that precise computational theories are also invaluable tools for understanding human learning.

7.1. Future Directions

In chapter 3 we mentioned that we avoided modeling children's preferences for *asking* specific informants for future information. The trust in testimony literature primarily works with word–learning paradigms which leaves the interpretation of children's preference unclear. Children may prefer to seek out informants from whom they have a better chance of learning or simply who are more likely to label correctly in the future. We would like to test whether children make a distinction between these strategies, whether they are context– and domain–specific; and whether children adopt different strategies at different ages. This is the pedagogical version of *active learning* (see Settles, 2009): learners do not choose their data, but they may choose the source of their data.

Additionally, we would like to expand the probabilistic model of learning from informants to account for more complicated epistemic phenomena. We could add power to our model by endowing it with a more robust prior. Certainly informants have variable knowledgeability and helpfulness but in the "real world" we must account for variance across situations and knowledge domains. For example, preschoolers understand that a car mechanic is more likely to know what cars are made of than a doctor (Lutz & Keil, 2002). It is not that the doctor is not knowledgeable; she simply has different expertise (is likely less knowledgeable about cars). Also, children see knowledgeability as domain–specific and ignorance as domain–general (Koenig & Jaswal, 2011). Research suggests that children's ability to monitor expertise is driven by their categorization abilities (Danovitch & Noles, 2014). We can capture a notion of expertise and potentially other information-seeking strategies by placing a cross–categorical prior on informants' helpfulness and knowledgeability. Under this prior, children can make inferences about informant's knowledgeability and helpfulness based on their categorization of knowledge domains (their perception of the structure of knowledge) and specific features of a given informant (e.g., an informant with the *has dogs* feature is likely to know about dogs).

7.2. Conclusion

We conclude by expressing our hopes for this work and for the future of the social learning literature. We hope that researchers will extend our model so that other social learning phenomena fall under its scope; that researchers will scrutinize our model and find and repair discrepancies; and that researchers will be motivated to reproduce others' work in neglected age and culture groups so that we may conduct more targeted meta-analyses. We hope that our work helps to focus the social learning literature and sparks a unified effort to reverse–engineer social learning.

REFERENCES

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: a psychological study of the strange situation*. Psychology Press.

Aldous, D. (1985). Exchangeability and related topics. *Ecole dÉté de Probabilités de Saint-Flour XIII*, *1117*, 1–198.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Asch, S. E. (1956). Studies of independence and conformity: i. a minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Aslin, R. N. & Newport, E. L. (2012, June). Statistical learning: From acquiring specific items to forming general rules. *Current directions in psychological science*, *21*(3), 170–176.

Bayes, T. & Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014, July). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive psychology*, *74C*, 35–65.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011, September). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–30.

Boyd, R., Richerson, P. J., & Henrich, J. (2011, June). The cultural niche: why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl*, 10918–25.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011, September). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–40.

Bulf, H., Johnson, S. P., & Valenza, E. (2011, October). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–32.

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002, May). What's new, pussycat? On talking to babies and animals. *Science (New York, N.Y.) 296*(5572), 1435.

Chen, E. E., Corriveau, K. H., & Harris, P. L. (2012). Children trust a consensus composed of outgroup members–but do not retain that trust. *Child development*, *84*(1), 269–82.

Collins, A. & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6).

Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009, March). Going with the flow: preschoolers prefer nondissenters as informants. *Psychological science*, *20*(3), 372–7.

Corriveau, K. H. & Harris, P. L. (2009, April). Choosing your informant: weighing familiarity and recent accuracy. *Developmental science*, *12*(3), 426–37.

Corriveau, K. H. & Harris, P. L. (2010, March). Preschoolers (sometimes) defer to the majority in making simple perceptual judgments. *Developmental psychology*, *46*(2), 437–45.

Corriveau, K. H., Harris, P. L., Meins, E., Fernyhough, C., Arnott, B., Elliott, L., . . . de Rosnay, M. (2009). Young Children's Trust in Their Mother's Claims: Longitudinal Links With Attachment Security in Infancy. *Child Development*, *80*(3), 750–761.

Couillard, N. & Woodward, A. (1999). Children's comprehension of deceptive points. *British Journal of Developmental Psychology*, *17*(4), 515–521.

Cristia, A. & Seidl, A. (2013). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, 1–22.

Csibra, G. (2008, May). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, *107*(2), 705–17.

Csibra, G. & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In *Processes of change in brain and cognitive development. attention and performance xxi* (pp. 249–274).

Csibra, G. & Gergely, G. (2009, April). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–53.

Csibra, G. & Volein,

bibinitperiod. (2008, March). Infants can infer the presence of hidden objects from referential gaze information. *British Journal of Developmental Psychology*, *26*(1), 1–11.

Danovitch, J. H. & Noles, N. S. (2014). Categorization Ability, but Not Theory of Mind, Contributes to Children's Developing Understanding of Expertise. *Proceedings of the 36th annual conference of the Cognitive Science Society.*

de Boer, B. & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129.

Dennett, D. C. (1978). Three kinds of intentional psychology. *Perspectives in the Philosophy of Language: A Concise Anthology*, 163–186.

DiYanni, C., Corriveau, K. H., Nasrini, J., Kurkul, K., & Nini, D. (under review). The role of consensus and culture in children's imitation of inefficient actions.

DiYanni, C. & Kelemen, D. (2008). Using a bad tool with good intention: young childre's imitation of adults' questionable choices. *Journal of experimental child psychology*, *101*(4), 241–261.

Eaves, B. S. & Shafto, P. (2012). Unifying pedagogical reasoning and epistemic trust. *Advances in child development and behavior*, *43*, 295–319.

Einav, S. & Robinson, E. J. (2010, July). Children's sensitivity to error magnitude when evaluating informants. *Cognitive Development*, *25*(3), 218–232.

Elashi, F. B. & Mills, C. M. (2014). Do children trust based on group membership or prior accuracy? the role of group membership in children's trust decisions. *Journal of Experimental Child Psychology.*

Feldman, J. (2000, October). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–3.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review*, *120*(4), 751.

Fink, D. (1997). A Compendium of Conjugate Priors. (1994), 1–47.

Fitneva, S. a. & Dunfield, K. a. (2010, September). Selective information seeking after a single encounter. *Developmental psychology*, *46*(5), 1380–4.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 721–741.

Gergely, G., Bekkering, H., & Király, I. (2002, February). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755.

Gergely, G. & Csibra, G. (2003, July). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gergely, G., Egyed, K., & Király, I. (2007, January). On pedagogy. *Developmental science*, *10*(1), 139–46.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008, September). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(37), 14222–7.

Gliga, T. & Csibra, G. (2009, March). One-year-old infants appreciate the referential nature of deictic gestures and words. *Psychological science*, *20*(3), 347–53.

Goldman, W. (2007). *The princess bride: s. morgenstern's classic tale of true love and high adventure*. Houghton Mifflin Harcourt.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007, December). Google and the mind: predicting fluency with PageRank. *Psychological science*, *18*(12), 1069–76.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Herbert, F. (2003). *Dune*. Penguin.

Hillenbrand, J., Getty, L. a., Clark, M. J., & Wheeler, K. (1995, May). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5 Pt 1), 3099–111.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*(1), 193–218.

James, W. (1890). *The principles of psychology*. Dover Publications.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, *186*(1007), 453–461.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007, May). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307–21.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012, January). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Kim, E., Song, G., Corriveau, K. H., & Harris, P. L. (2013, January). Young Children's Deference to a Consensus Varies by Culture and Judgment Setting. *Journal of Cognition and Culture*, *13*(3-4), 367–381.

Kirchhoff, K. & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002, March). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–42.

Koenig, M. & Echols, C. (2003). Infants' understanding of false labeling events: the referential roles of words and the speakers who use them. *Cognition*, *87*, 179–208.

Koenig, M. & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child development*, *76*(6), 1261–77.

Koenig, M. & Jaswal, V. K. (2011). Characterizing children's expectations about expertise and incompetence: halo or pitchfork effects? *Child development*, *82*(5), 1634–47.

Kondrad, R. L. & Jaswal, V. K. (2012, April). Explaining the errors away: Young children forgive understandable semantic mistakes. *Cognitive Development*, *27*(2), 126–135.

Kuhl, P. K., Andruski, J. E., Christovich, I. A., Christovich, L. A., Kozhevinkova, E. V., Ryskina, V. L., … Lacerda, F. (1997, August). Cross-Language Analysis of Phonetic Units in Language Addressed to Infants. *Science*, *277*(5326), 684–686.

Lutz, D. J. & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child development*, *73*(4), 1073–84.

MacEachern, S. N. & Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of computational and graphical statistics*, *7*(2).

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. The MIT Press.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009, April). Statistical learning of phonetic categories: insights from a computational approach. *Developmental science*, *12*(3), 369–78.

McMurray, B., Kovack-Lesh, K. a., Goodwin, D., & McEchron, W. (2013, November). Infant directed speech and the development of speech perception: enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–78.

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology*, *24*(4), 470–476.

Meltzoff, A. N. (1995). What infant memory tells us about infantile amnesia: Long-term recall and deferred imitation. *Journal of experimental child psychology*, *59*, 497–515.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution. *Technical Report*.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, *9*(2), 249–265.

Neapolitan, R. E. et al. (2004). *Learning bayesian networks*. Prentice Hall Upper Saddle River.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web., 1–17.

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007, September). Preschoolers monitor the relative accuracy of informants. *Developmental psychology*, *43*(5), 1216–26.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ Press.

Piaget, J. (1999). *Construction of reality in the child*. Psychology Press.

Piantadosi, S. T., Kidd, C., & Aslin, R. (2014, February). Rich analysis and rational models: inferring individual behavior from infant looking data. *Developmental Science*.

Pinker, S. (2010, May). The cognitive niche: coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences of the United States of America*, *107 Suppl*, 8993–9.

Rasmussen, C. (2000). The infinite Gaussian mixture model. *Advances in neural information processing*, (11), 554–560.

Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. In *Combinatorial mathematics v* (pp. 28–43). Springer.

Russell, S. & Norvig, P. (2009). *Artificial intelligence: a modern approach* (3rd ed.). Prentice Hall.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical Learning by 8-month-Old Infants. *Science*, *274*(5294), 1926–1928.

Scott, R. M. & Baillargeon, R. (2013, April). Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychological science*, *24*(4), 466–74.

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Shafto, P., Eaves, B. S., Navarro, D. J., & Perfors, A. (2012, May). Epistemic trust: modeling children's reasoning about others' knowledge and intent. *Developmental science*, *15*(3), 436–47.

Shafto, P. & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the thirtieth annual conference of the cognitive science society*.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014, March). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71C*, 55–89.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011, July). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1–25.

Shepard, R. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323.

Sloman, S., Love, B., & Ahn, W. (1998, April). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search.* Lecture Notes in Statistics. New York, NY: Springer New York.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006, December). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Tenenbaum, J. B. & Griffiths, T. L. (2001). The rational basis of representativeness. *. . . of the 23rd annual conference of the . . .*

Tomasello, M. (2009). *The cultural origins of human cognition.* Harvard University Press.

Topal, J., Gergely, G., Miklosi, A., Erdohegyi, A., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science*, *321*(September), 1831.

Tummeltshammer, K. S., Wu, R., Sobel, D. M., & Kirkham, N. Z. (2014, July). Infants Track the Reliability of Potential Informants. *Psychological science*, (July).

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007, August). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(33), 13273–8.

A. MONTE CARLO ESTIMATION IN THE EPISTEMIC TRUST MODEL

A.1. Rejection sampling

Rejection sampling[*] is a Monte Carlo method in which individual samples are independently generated from a model and accepted (collected or counted) if they meet some criteria and rejected otherwise. To draw a sample:

1. If w is not known, draw w from $P(w)$
2. Draw θ_k from $beta(\alpha_k, \beta_k)$ and draw θ_h from $beta(\alpha_h, \beta_h)$
3. Draw k from Bernoulli(θ_k) and draw h from Bernoulli(θ_h)
4. Draw b from $P(b|k, w)$
5. Draw a from $P(a|h, b)$
6. Draw e from $P(e|a, w)$

We can estimate probabilities by comparing the frequency with which certain variables in the sample match observed values.

For example, if we wished to calculate the probability that an informant was knowledgeable and not helpful given that she produced a certain label, say $l = 1$, on an accuracy trial that did not match the state of the world, say $w = 2$ we would count the number of samples in which k, $\neg h$, $l = 1$, and $w = 2$ are true and divide that by the number of samples in which $l = 1$, and $w = 2$ are true. This is an estimate of the conditional probability,

[*] For a review of sampling methods in Bayesian networks see Russell and Norvig (2009), Neapolitan et al. (2004)

$$P(k, \neg h | l = 1, w = 2) = \frac{P(k, \neg h, l = 1, w = 2)}{P(l = 1, w = 2)} \approx \frac{N(k, \neg h, l = 1, w = 2)}{N(l = 1, w = 2)}, \tag{A.1}$$

where $N(\cdot)$ is the number of samples in which the each condition within is met.

We can also estimate continuous distributions using rejection sampling by storing the values of continuous parameters. For example, during accuracy trials we wish to estimate the distribution of each informant's θ's. If the observed label is l_o and the observed world state is w_o, we collect θ_{k_i}, the θ_k value on the ith sample and add it to the bin if $w_i = w_o$ and $l_i = l_o$. We do similarly with θ_h. We calculate ask probability, $P(\text{ask})$, using the values in these vectors also using rejection sampling. We generate a sample from the model using the values in $\theta_{\mathbf{k}}$ and $\theta_{\mathbf{h}}$ rather than drawing from the beta prior distribution. We simply divide the number of samples for which $l = w$ by the total number of samples.

Rejection sampling is accurate but inefficient. Though each sample is guaranteed to be independent because each sample is drawn independently, many of the generated samples are unused. The lower the probability of the event conditioned on, the fewer samples are collected. Rejection sampling in the epistemic trust model is prohibitively slow over multiple trials and multiple informants.

A.2. Gibbs sampling

Gibbs sampling (see S. Geman & D. Geman, 1984; Gelman et al., 2013) does not waste samples and is well-suited for use in Bayesian networks. It works by re-sampling each node conditioned on the values of every other node in the network. However, it is most often the case that a node is *not* dependent on every other node but only a few. The terms in which the target node does not appear cancel out. We can exploit conditional dependence to simplify further. A node is conditionally independent of all other nodes in a network given its *Markov blanket*: the nodes comprising its parents, children, and children's parents. The Markov blanket of the knowledgeability node can be seen shaded in gray in Figure A.1. The conditional probabilities and distributions of each variable are thus:

$$
\begin{aligned}
\theta_k &\sim \text{beta}(\alpha_k + n_k, \beta_k + n_{\neg k}), && \text{(A.2)}\\
\theta_h &\sim \text{beta}(\alpha_h + n_h, \beta_h + n_{\neg h}), && \text{(A.3)}\\
w &\sim p(w)p(b|k,w)p(e|a,w), && \text{(A.4)}\\
k &\sim p(k|\theta_k)p(b|k,w), && \text{(A.5)}\\
h &\sim p(h|\theta_h)p(a|h,b), && \text{(A.6)}\\
b &\sim p(b|k,w)p(a|h,b), && \text{(A.7)}\\
a &\sim p(a|h,b)p(e|a,w), && \text{(A.8)}\\
e &\sim p(e|a,w), && \text{(A.9)}
\end{aligned}
$$

where n_h and $n_{\neg h}$ are the number of trials in which the informant has been helpful and unhelpful, and where n_k and $n_{\neg k}$ are the number of trials in which the informant has been knowledgeable and unknowledgeable.

The sampler state is set to some random value fixing observed nodes to their observed values. For a predetermined number of iterations, the Gibbs sampler updates each unobserved node in random order. For example, if we observe an action and an effect, we set the a and e nodes and update all other nodes while keeping a and e static. We then collect or count as we did with rejection sampling subject to some caveats.

Samples generated by a Gibbs sampling algorithm are not independent. They depend on the previous state. To mitigate effects of sample interdependence we ignore a certain number of samples between each collection. This process is known as *lag* or *thinning.* For the same reason, we must ignore some number of samples before collecting the first. The sampler state may have been initialized to a value that is not representative of the target distribution and it make take the sampler some time to walk its way to the target region. Another concern is Gibb samplers' propensity to get stuck in local maxima. Imagine a bimodal probability distribution with two distant peaks. In order for the sampler to cross the gap from peak to peak it must cross a large space of low probability. It is common practice to average samples over multiple independent instances (*chains*) of Gibbs sampler

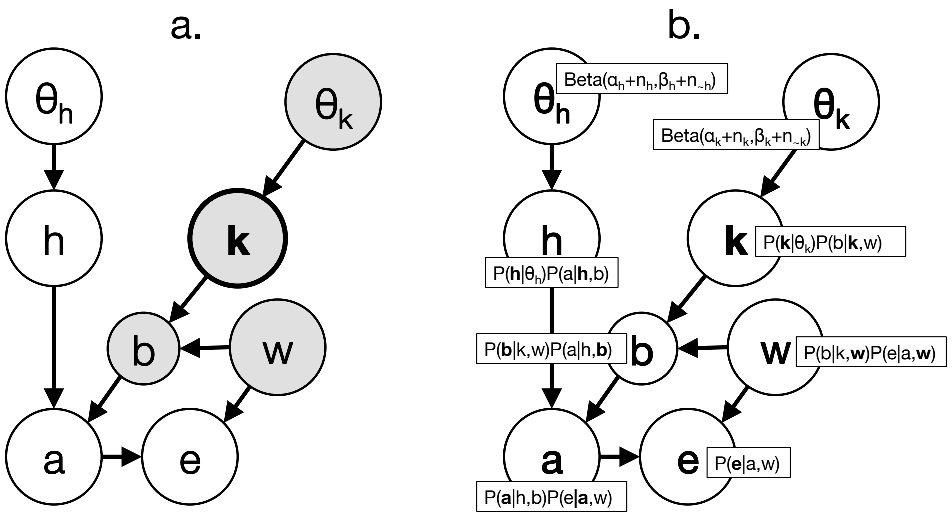


Figure A.1: Example Markov blanket for epistemic trust model and Gibbs sampling conditional probabilities. a) Markov blanket (shaded in gray) for knowledgeability, k. b) Gibbs sampling conditional probabilities and distribution superimposed on their respective variables.

runs to smooth the between–chain variability due to local maxima.

B. EXPECTED DEPENDENCE PROBABILITY BETWEEN CROSS-CATEGORIZATION COLUMNS

First, we derive the probability, under the Chinese Restaurant Process (CRP), that two items will be assigned to the same component. Because the CRP is an exchangeable process, in the limit it may be described as i.i.d. This means that we need only be concerned with the probability that the *first two* items are assigned to the same component. The first item is always assigned to its own component; the second item is assigned the the same component with probability $\frac{1}{1+\alpha}$, where α is the CRP concentration parameter. Thus the probability that any two columns, i and j, belong to the same components is,

$$P(z_i = z_j|\alpha) = \frac{1}{1+\alpha}. \tag{B.1}$$

In our implementation of cross-categorization α is given an exponential prior with mean 1. That is,

$$\alpha \sim \text{Exp}(1). \tag{B.2}$$

We must calculate the expected dependence probability across the prior. That is,

$$E[Y] = E\left[\frac{1}{1+\alpha}\right]. \tag{B.3}$$

We derive the cdf of this distribution:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) && \text{(B.4)} \\ &= P\left(\frac{1}{y} - 1 \leq \alpha\right) && \text{(B.5)} \\ &= 1 - F_X\left(\frac{1}{y} - 1\right). && \text{(B.6)} \\ & && \text{(B.7)} \end{aligned}$$

Differentiating leaves us with the pdf:

$$f_Y(y) = \frac{1}{y^2}\exp\left(1 - \frac{1}{y}\right). \tag{B.8}$$

The expected dependence probability between two columns is

$$P(z_i = z_j) = E\left[Y\right] = \int_0^1 y f_Y(y) dy = \int_0^1 \frac{1}{y}\exp\left(1 - \frac{1}{y}\right) dy \approx 0.596. \tag{B.9}$$

CURRICULUM VITAE

Baxter Eaves

Department of Psychological and Brain Sciences
317 Life Sciences Building
University of Louisville
Louisville, KY 40292
Baxter.Eaves@louisville.edu

Born: May 16, 1984—Miami, Florida

Education

2014 Ph.D. Experimental Psychology
University of Louisville

2013 M.S. Experimental Psychology
University of Louisville

2011 B.A. Psychology
University of Louisville

Awards

2014 Graduate Dean's Citation

Publications

Eaves, B., Feldman, N., Giffiths, T., & Shafto, P. (under review). Infant-directed speech as optimal input for learning vowel categories . Cognition.

Landrum, A., Eaves, B., & Shafto, P. (under review). Trusting to learn and learning to trust. Trends in Cognitive Science.

Eaves, B. & Shafto, P. (2014). Order effects in learning relational structures. Proceedings of the 36th annual conference of the Cognitive Science Society.

Eaves, B. & Shafto, P. (2012). Unifying pedagogical reasoning and epistemic trust. In Xu. F. and Kushnir, T(eds.) Advances in Child Development and Behavior.

Shafto, P., Eaves, B., Navarro, D.J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others knowledge and intent. Developmental Science.

Grants and Fellowships

University of Louisville Graduate Research Fellowship

University of Louisville Undergraduate Research Grant

University of Louisville Undergraduate Research Initiation Grant

Posters and Presentations

2014 Order effects in learning relational structures. Proceedings of the 36th annual conference of the Cognitive Science Society.

2013 Infant-directed speech as statistically optimal input. Eighth Biennial Meeting of the Cognitive Development Society

2012 Learning from others: A computational model of epistemic trust. Midwest Cognitive Science Conference

2012 Learning from others: A computational model of epistemic trust. University of Louisville Graduate Research Symposium

2012 Modeling Epistemic Trust and Implications for Learning. Society for Research Child Development

Open Source Software Contributions

BayesDB—A Bayesian database that lets users query the probable implications of their data with a SQL-based query language

CrossCat—A domain-general, hierarchical Bayesian non-parametric method for analyzing high-dimensional data tables.