Causal Models in Prediction and Diagnosis

By Philip M. Fernbach

B. A., Williams College, 2001

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Department of Cognitive and Linguistic Sciences at Brown University

Providence, Rhode Island

May 2010

This dissertation by Philip M. Fernbach is accepted in its present form by the Department of Cognitive and Linguistic Sciences as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Dr. Steven Sloman, Advisor

Recommended to the Graduate Council

Date _____ _____

Dr. Bertram Malle, Reader

Date _____ _____

Dr. David Sobel, Reader

Approved by the Graduate Council

Date _____ _____

Dr. Sheila Bonde, Dean of the Graduate School

VITA

Philip M. Fernbach

Department of Cognitive & Linguistic Sciences
Brown University, Box 1978
Providence, Rhode Island 02912 USA
(401) 863-1167

200 Governor St. unit C
Providence, Rhode Island 02906
(401) 383-5361

email: philip_fernbach@brown.edu
website: http://sites.google.com/site/philfernbachswebpage/

Born: May 24, 1979, Albany, NY, USA

Education

Doctoral Student in Cognitive Science, Brown University, Providence, RI, September 2005-Present, PhD expected Spring 2010

Graduate Summer School: Probabilistic Models of Cognition: The Mathematics of Mind, UCLA Institute for Pure and Applied Mathematics, Los Angeles, CA, Summer 2007

B.A., Philosophy and Pre-Medicine, Williams College, Williamstown, MA, June 2001.

Employment

2003-2005, Analyst, Auctive Incorporated, Boston, MA: Headed the research department. Designed and executed research projects for strategy consulting clients in the consumer packaged goods industry.

2002-2003, Research Analyst, Dove Consulting, Boston, MA: Worked as part of a team on strategy and management consulting projects for clients in the beverage and food, consumer broadband and financial services industries. Primary responsibilities included conducting research and analyzing data, and financial modeling.

Honors and Awards

American Psychological Association Dissertation Research Award, 2009

Galner Dissertation Fellowship, Brown University, 2009-2010

Funded Participant: Graduate Summer School: Probabilistic Models of Cognition: The Mathematics of Mind, UCLA Institute for Pure and Applied Mathematics, 2007

Cognitive Science Society Student Travel Award, 2007, 2009

Brown University Graduate Fellowship, 2005-2006

Magna Cum Laude, Williams College, 2001

Refereed Publications

Fernbach, P. M. & Darlow, A. (2010). Causal conditional reasoning and conditional likelihood. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

Sloman, S. A., Fernbach, P. M. & Hagmayer, Y. (in press). Self-deception requires vagueness. *Cognition*.

Fernbach, P. M., Darlow A. & Sloman, S. A. (in press). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*.

Fernbach, P. M. & Darlow, A. (2009). Causal asymmetry in inductive judgments. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

Fernbach, P. M. & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35 (3), 678-693.

Fernbach, P. M., Linson-Gentry, P & Sloman, S. A. (2007), Causal beliefs influence the perception of temporal order (submitted). *Proceedings of the twenty-ninth Annual Conference of the Cognitive Science Society*.

Fernbach, P. M. (2006). Sampling assumptions and the size principle in property induction. *Proceedings of the twenty-eighth Annual Conference of the Cognitive Science Society*.

Book Chapters

Sloman, S. A., Fernbach, P. M. & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In D. Bartels, C. W. Bauman, L. J. Skitka, & D. Medin (Eds.) *Moral judgment and decision-making: The psychology of learning and motivation (Vol. 50)*. San Diego, CA: Elsevier.

Sloman, S. A. & Fernbach, P. M. (2008). The value of rational analysis: An assessment of causal reasoning and learning. In Chater, N. & Oaksford, M. (Eds.). *The probabilistic mind: Prospects for rational models of cognition*. Oxford: Oxford University Press.

Submitted Journal Articles

Fernbach, P. M., Darlow A. & Sloman, S. A. (under review). Asymmetries in predictive and diagnostic reasoning.

Sloman, S. A., Fernbach, P. M. & Ewing, S. (under review). A Causal model of intentionality judgment.

Conference and Workshop Presentations

Fernbach, P. M. & Darlow A. (August, 2009). *Causal asymmetry in inductive judgments*. 31st annual meeting of the Cognitive Science Society, Amsterdam, Netherlands.

Fernbach, P. M. (March, 2008). *Causal Reasoning in the Wild*. Meeting of the Eastern Psychological Association, Boston, MA.

Fernbach, P. M., Linson-Gentry, P & Sloman, S. A. (August, 2007). *Causal beliefs influence the perception of temporal order*. 29th annual meeting of the Cognitive Science Society, Nashville, TN.

Fernbach, P. M. (September, 2006). *Heuristic causal learning*. Brown University Causality Workshop, Providence, RI.

Fernbach, P. M. (August, 2006). *Sampling assumptions and the size principle in property*, 28th annual meeting of the Cognitive Science Society, Vancouver, Canada.

Service

Founder and organizer of the Computational Modeling Reading Group, 2007-present; Website: http://sites.google.com/site/philfernbachswebpage/cmrg

Teaching

Spring, 2008: Teaching Assistant, CG 105: Music and the Mind

Fall, 2008: Teaching Assistant, CG 109: Quantitative Methods in Psychology

Spring, 2007: Teaching Assistant, CG 152: Thinking

Fall, 2007: Teaching Assistant, CG 001: Introduction to Cognitive Science

Fall, 2006: Teaching Assistant, CG 153: Laboratory in Cognitive Processes

Reviewing

Cognition
Cognitive Science
European Journal of Cognitive Psychology
Experimental Brain Research
Journal of Problem Solving
Memory and Cognition
Psychological Science
Quarterly Journal of Experimental Psychology
Conference of the Cognitive Science Society; 2007-2010
Conference of the Eastern Psychological Association; 2008

Preface and Acknowledgments

The work reported in this dissertation would not have been possible without the contributions of many people. Several people made contributions to individual portions of the work, and are acknowledged in the notes at the end of the document. I'd like to thank a few people for overall contributions. First, thanks to Dave Sobel and Bertram Malle for serving on my committee, and offering constructive feedback and many valuable discussions of work and life. Adam Darlow was a collaborator and co-author on several publications emerging from the dissertation. He made important contributions to many aspects of the work including in the development and testing of the model. He was a wonderful collaborator and good friend, and I thank him for that. Steve Sloman, my dissertation advisor has been a wonderful mentor and friend to me throughout graduate school. He is also a co-author on several publications emerging from the work described here and made valuable contributions in many areas. Finally, I could not have completed this work without the support of my amazing family: Joan, Alex and Rachel Fernbach; and Bruce and Joyce Slater. Most of all I'd like to thank my wife and love of my life Anna for always being there for me.

Table of Contents

List of Tables

List of Figures

Chapter 1:Introduction

Chapter 2: Asymmetries in Predictive and Diagnostic Reasoning

Chapter 3: Neglect of Alternative Causes

Chapter 5: Reaction Times and Causal Conditional Reasoning

1. Introduction

Our decisions are often guided by quick and intuitive assessments of likelihood. You might look out the window, predict that rain is likely, and decide to pack an umbrella. You might notice a broken vase in the living room, infer your son broke it and decide to punish him. The facility with which we make such inferences belies how much goes into making them. Given how natural they seem, it is tempting to assume that they must rely on some simple set of rules or associations. As I will argue, such judgments actually require sophisticated representational capacities.

Consider the two questions below:

a) He has a strong motive. How likely is it he committed a crime?

b) He's acting suspiciously. How likely is it he committed a crime?

In many ways, they are quite similar. In both cases, a state of affairs (a motive or suspicious behavior) provides evidence for an unknown conclusion (committing a crime), the inference is uncertain, the syntax is almost identical, and so on. However, there is an important difference; the two inferences differ in the direction of the causal relation between evidence and conclusion. A motive causes the commission of a crime whereas suspicious behavior is an effect of the commission of a crime. Throughout the dissertation I refer to inferences from causes to effects as *predictions* and inferences from effects to causes as *diagnoses*. As I will show, causal directionality has profound and systematic effects on judgment. The studies reported in this dissertation identify several empirical phenomena that cast light on the processes and representations that support predictive and diagnostic reasoning. For instance, when making predictions, people fail to think about alternative causes and instead focus on how likely a given mechanism is to

lead to an effect. When making a diagnosis they think more broadly, considering alternative causes and adjusting their judgments accordingly. Diagnosis is also slower, more difficult, and develops later than prediction. The purpose of the studies described in this dissertation is to elucidate the role that causal directionality plays in people's intuitive assessments of likelihood with an eye to explaining these phenomena.

1.1 Causal Models

Should causal directionality matter? As Tversky and Kahneman (1982) write, "In a normative treatment of conditional probability the distinction between the various types of relations … are immaterial, and the impact of data depends solely on their informativeness" (p 118). They go on to argue that causal directionality biases judgment because people find it easier to reason from causes to effects than vice-versa and therefore overweight the predictive direction in judgment (the 'causal asymmetry conjecture'). The upshot is that causal directionality is a nuisance that gets in the way of what would otherwise be more accurate judgment.

A different possibility is that causal reasoning, far from being detrimental, is instrumental to good judgment (Nozick, 1993). While it is true that given a certain evidential impact, the causal role of data should not matter, it is not typically the case that informativeness is self-evident. Instead it has to be inferred by piecing together the relevant information. Causal roles constrain what is relevant and hence can inform the retrieval of additional necessary information. Consider b) above. Making this judgment requires thinking about other reasons why he might be acting suspiciously. Such information is not present in the immediate discourse context, but the causal role of the evidence indicates that it is important. The likelihood of a cause given an effect depends

on the presence of other causes. To the extent that other good explanations for his suspicious behavior are present, the conclusion becomes less likely. The same logic does not apply to the predictive judgment.

I began this work with the hypothesis that people's facility for intuitive likelihood assessment is based in their ability to faithfully represent key aspects of the world's causal structure and to derive sound judgments from that representation. I refer to this idea as the 'causal model conjecture.' This conjecture follows from a growing body of evidence speaking to the centrality of causal reasoning in human cognition. For instance, people reason more naturally about causal systems than other kinds of systems (Bindra, Clarke & Schultz, 1980), their judgments are well described by causal logic (Sloman & Lagnado, 2005) and their concepts are organized according to causal structure (Reheder, 2003). Theories aimed at accounting for these and other consistent findings are usually called 'causal model theories' (Sloman, 2005, Gopnik et al. 2004, Glymour, 2001). According to causal model theories, mental representations are causal in nature as opposed to associative or logical and the goal of learning is to recover an accurate model of the world's causal structure that can then be used to make uncertain inferences (Waldmann & Holyoak, 1992).

To make this more concrete, consider the predictive question in (c) and the diagnostic one in (d).

(c) A mother has a drug-addiction. How likely is it that her newborn baby has a drug-addiction?

(d) A newborn baby has a drug-addiction. How likely is it that the baby's mother has a drug-addiction?

A causal model theory would suggest that people make these judgments by building or retrieving a causal model of the scenario, which includes the information necessary for computing the desired quantity. This includes knowledge about causal structure (e.g. that a mother's drug addiction is a cause of the baby's drug addiction, and not vice-versa), strength of causal relations, potential alternative causes, disabling conditions, base rates and so on. A reasonable model might look something like Figure 1.1. Once the model is in place, evaluating c) and d) requires computing likelihood from the primitives that make up the model in a way that yields a reliable estimate of the true probability distributions.

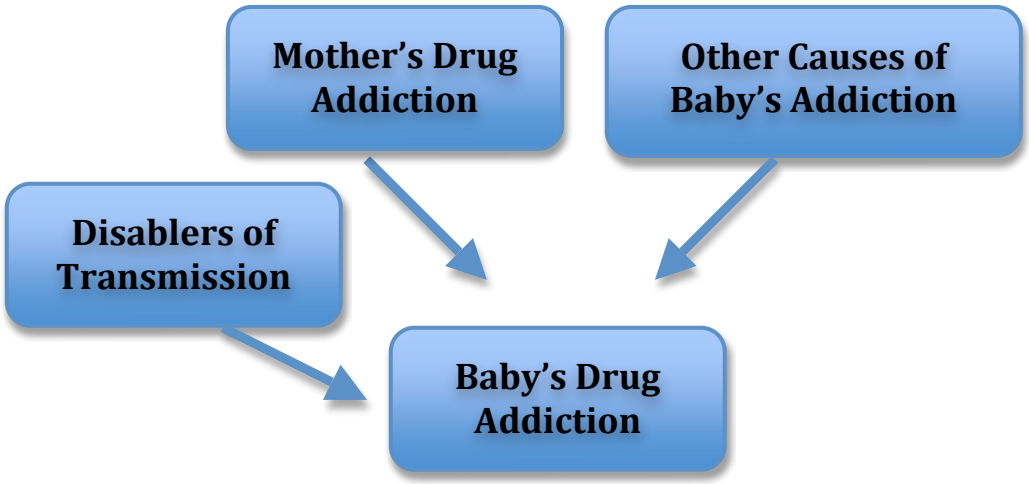

Figure 1.1 A simple causal model of transmission of a drug-addiction between mother and baby

The idea that such a model building and evaluation process occurs during intuitive likelihood assessment is a departure from other popular theories of judgment. One prevailing approach to subjective likelihood judgment theorizes that judgments depend on a small number of simple evaluative heuristics, like counting available instances in memory or comparing the similarity between evidence and conclusion (Kahneman, Slovic & Tversky, 1982). A different approach views likelihood judgment as a process of

'naïve extensional reasoning' whereby people enumerate the possible outcomes and distribute belief over them uniformly (Fox & Levav, 2004, Johnson-Laird at al, 1999). In neither of these approaches does causal directionality play a central role. In contrast, the causal model conjecture implies a sophisticated causal reasoning process that involves integrating various pieces of information into judgment depending on the causal relations connecting evidence and conclusion.

If people are using causal models in their predictive and diagnostic judgments, then judgment should differ systematically and predictably based on causal directionality. But how should it differ? Answering this question requires a normative model that specifies what factors matter and precisely how they interact to determine predictive and diagnostic likelihood. Such a model is developed in Chapter 2 so will not be described in detail here. Instead I will simply highlight the key factors to give an intuition for the predictions of the causal model conjecture. The first factor is 'causal power'. Causal power is the likelihood that a causal mechanism, when present successfully brings about its effect (Cheng, 1997). In the example above, the causal power refers to the likelihood that the mother transmits her addiction to her baby. Predictive and diagnostic judgments should both decrease as causal power decreases. Weak causes are less likely to generate their effects and effects, when present, are less likely to be due to weak causes. The second factor is the strength of alternative causes. This refers to the prevalence of other causal routes by which the effect could happen, for instance other ways that a baby could become drug-addicted. This factor has an opposite effect on prediction and diagnosis. As the strength of alternative causes increase, predictive judgment should increase because the probability of the effect is higher, but diagnostic judgment should decrease because

alternative causes provide an alternative explanation for the effect. Finally, the prior probability of the cause is relevant to diagnostic judgment only. All else being equal, if a cause is rare, it provides a poorer explanation for the effect as compared to alternative causes. There is a greater chance that the cause was absent and the effect was due to some other cause.

The causal analysis yields a set of predictions as to how predictive and diagnostic judgments should change as a function of causal directionality and the underlying causal beliefs about the scenario being judged, if the causal model conjecture is correct. The first set of studies (described in Chapter 2) was designed to test these predictions. Broadly, two findings were noteworthy and set the stage for the remaining lines of work in the dissertation.

First, the causal model conjecture obtained strong support. People brought their background knowledge to bear on predictive and diagnostic judgments, without instruction, and they were sensitive to the relevant factors in approximately the right ways. Diagnoses were primarily influenced by the strength of alternative causes and the base rate of the focal cause. Predictions were primarily based on the causal strength of the focal cause. The quantitative model was highly correlated with both predictive and diagnostic judgments with zero free parameters. These findings support the hypothesis that people make predictive and diagnostic judgments by deriving them from a reasonable, self-generated causal model. Taken on its own this outcome was informative as to the key question posed above: what is the role of causal directionality in intuitive likelihood assessment? At least part of the answer is that the causal role of evidence

determines its place in a causal model of the scenario being judged and hence, aids people in assessing its informativeness.

The second notable finding is that there was one source of systematic bias. People failed to think about alternative causes when making predictions and therefore underestimated the likelihood of effects given causes. As described above, alternative causes are important to prediction because they raise the probability of the effect. For instance, in predicting whether someone who lives near a power plant will get cancer, it is important to consider whether he or she is also a smoker. This error provided a hook for elucidating the underlying process and it inspired much of the rest of the work reported in the dissertation.

1.2 Causal Myopia

What explains such an effect? I argue that when possible, people prefer to think about one causal mechanism at a time and they focus on whichever one is most easily retrieved. Therefore the cause one is conditioning on dominates thought to the detriment of relevant alternatives, a tendency I refer to as 'causal myopia.' This idea is inspired by related findings from the heuristics and biases literature. Work in this tradition is primarily aimed at understanding when and why people's intuitive likelihood assessments depart from norms. A unifying approach to explain these errors, proposed by Tversky and Koehler (1994) is referred to as Support Theory. According to the theory, subjective likelihood judgments are tied to descriptions of events and not to the events themselves. Since the way an event is described can alter how we think about it, different descriptions can lead to different judgments. The theory is referred to as a 'non-extensional' theory of belief because such effects violate the extensionality principal of probability theory, that the

same event must have the same probability no matter how it is described. Evidence for the theory comes primarily from studies showing differing likelihood judgments depending on how an event is described. A prototypical case occurs when a hypothesis is unpacked into its constituents. For instance, participants judge 'homicide' to be less likely than 'homicide by an acquaintance or homicide by a stranger' (Rottenstreich & Tversky, 1997). While the two descriptions refer to the same event (any case of homicide) unpacking the event into constituents evidently causes people to think about the problem differently. According to Support Theory, adding detail to the description increases support for the hypothesis and increases likelihood judgments. Why does adding detail increase support? One contributing factor is availability. People tend to base judgments on a small sample of memory, and the content of this sample depends on the ease of retrieval (Tversky & Kahneman, 1973). Examples that are more easily retrieved are more likely to affect judgment. In the example above, unpacking homicide into constituents may remind people of cases of homicide that they would not have thought of otherwise. To the extent that the sample is unrepresentative or incomplete, errors may occur. This tendency to base judgment on only a subset of the necessary information is a hallmark of human cognition that appears across a wide variety of tasks, and I review some examples in chapter 2.

Like with unpacking effects, the neglect of alternative causes in prediction can lead to violations of extensionality. For instance, (e) was judged greater than (f) (Experiments 1 and 3 in Chapter 2).

(e) The coach of a high school football team is highly motivated. Accolades from family and friends could also cause high school football teams to be highly motivated. How likely is it that the team is highly motivated?

(f) The coach of a high school football team is highly motivated. How likely is it that the team is highly motivated?"

The extensions are the same because no additional information is given in (e). It is generally known that accolades from family and friends can motivate teams. This statement only reminds participants of something they already know. The difference in judgment is due to the neglect of alternative causes in (f). When answering that question people think of just the single most available cause (the coach) but when they are reminded of an additional cause, they take it into account and raise the judgment accordingly.

Neglect of alternative causes can even lead to more extreme errors. Chapter 4 describes a series of experiments identifying a non-monotonic effect in predictive judgment, the weak evidence effect. When asked to judge the likelihood of an effect given a weak cause, participants actually gave lower judgments than when asked to judge the marginal likelihood of the effect. For example, (e) was judged *lower* than (f) even though the power outage was judged to raise the likelihood of spoilage in a separate condition (Experiment 2 in Chapter 4).

(e) A man buys a half-gallon of milk on Monday. The power goes out for 30 minutes on Tuesday. How likely is it the milk is spoiled a week from Wednesday?

(f) A man buys a half-gallon of milk on Monday. How likely is it the milk is spoiled a week from Wednesday?

When a weak cause is mentioned, participants focus on it and fail to consider alternative, stronger causes. When the weak cause is not mentioned they estimate the likelihood of the effect by retrieving stronger, more typical causes.

1.3 Theoretical Implications

These findings prompt a key question: If the availability of the focal causal mechanism biases people to ignore alternative causes in prediction, why are they sensitive to alternative causes when making diagnoses? Part of the answer is teleological. As will be explored in greater detail in Chapter 2, prediction and diagnosis differ in the normative role that alternative causes play in determining conditional likelihood. Diagnostic inference is intrinsically comparative in that diagnostic likelihood is, in part, a measure of how good an explanation the focal cause is *relative* to other causes. An estimate of diagnostic likelihood that ignored alternative causes would be ecologically poor, uncorrelated with the true probability distribution. Conversely, focusing on a single cause and its causal power can often be a good strategy for prediction, especially when the cause is relatively strong, since the magnitude of error is bounded by the difference between the causal power of the focal cause and one.

Heuristics are valuable because they minimize effort while yielding reasonable estimates most of the time. For instance, availability is a good strategy because searching through memory for relevant information is time- and energy-consuming. No judge can be expected to think about everything that is relevant to a given problem. When making a prediction or a diagnosis, the focal cause provides a starting point. Judging causal

power can be accomplished by focusing, at least temporarily, on that causal mechanism and thinking about whether it is likely to lead to the effect. Thinking about alternatives, however requires going beyond the context of the question and making a 'global' inference. It seems that people are not willing to engage in such a cognitive process unless it is strictly necessary.

These ideas suggest that prediction seems easier or more natural, not because people use causal schemas in a biased way as suggested by Tversky and Kahneman (1982), but rather because diagnostic inference prompts a different kind of thinking that is more effortful. One piece of evidence in support of this idea is that contrary to the causal asymmetry conjecture, which predicts a bias for predictive judgments to be too high relative to diagnostic judgments, the evidence reported in this dissertation shows a different pattern; diagnostic judgments were relatively unbiased while predictions were too low due to the neglect of alternative causes. As described below, two additional sources of evidence for different processing in the two directions of reasoning comes from studies of reaction time and development.

We are now in a position to provide a more complete answer to the role of causal directionality in intuitive assessments of likelihood. The answer has two aspects. First, causal directionality influences judgment by determining the role of the evidence in the causal model of the scenario being judged. People do not simply judge the conditional likelihood of conclusion given evidence. They judge the likelihood of effects given causes and causes given effects, and they do so by generating and evaluating a causal model that is mostly faithful to the system it is meant to represent. Second, people prefer to make predictions by focusing on an individual causal mechanism and thinking forward

to likelihood of the effect given that cause, but they think more broadly when making diagnoses, considering the relative explanatory power of relevant alternative causes. Thus, causal directionality also influences intuitive likelihood assessment in determining the kind of cognitive processes that are used for judgment.

1.4. Dissertation Roadmap

The dissertation is comprised of 5 lines of work. The first line (described in Chapter 2) tests the causal model conjecture in the domain of property transmission arguments like the drug-addiction example above (cases where a property is transmitted between two categories). A generic causal model for transmission scenarios is developed and equations for predictive and diagnostic conditional probability are derived from the model to serve as a normative benchmark. In the experiments, I collect judgments of the primitives of the model (prior probability, causal power and strength of alternatives) and derive model predictions. The model predicts how predictive and diagnostic judgments should vary as function of the underlying beliefs. Evidence for the causal model conjecture comes from comparing the model fits to the judgments.

As described above, the model fitting provided indirect evidence that people neglect alternative causes in prediction but not diagnosis. Chapter 3 corroborates this phenomenon with a more direct manipulation. Three experiments compare predictive and diagnostic judgments about 'full' and 'no-alternative' conditional likelihoods. Full conditionals are standard conditional likelihood questions. No-alternative conditionals rule out all causes other than the one mentioned. If people neglect alternative causes, there should be no difference between responses to the two types of conditionals. The goal of Chapter 4 is test the boundary conditions for the neglect of alternative causes in

prediction. The magnitude of error due to neglect of alternative causes increases as the causal power of the focal cause decreases. An extreme case occurs when the presence of a weak cause leads to lower judgments than when no cause is mentioned. I refer to this as the 'weak evidence effect'.

If diagnosis and prediction rely on different cognitive processes then such differences should be reflected in the amount of time it takes to answer questions. Chapter 5 tests this hypothesis by collecting reaction times for predictive and diagnostic judgments about causal scenarios adopted from Cummins (1995). The most basic hypothesis is that diagnosis should be slower than prediction because of the presence of a search process for alternative causes. Secondary hypotheses are that reaction time for diagnostic judgments should vary with the strength of alternatives and that predictive judgments should vary with causal power.

There is some evidence that reasoning forward from causes to effects develops earlier than the reasoning from effects to causes (Bindra, Clark & Shultz, 1980; Hong et al. 2004). If prediction is more natural than diagnosis because of the absence of a process for searching for and comparing alternatives to the focal cause, this could explain such findings. Chapter 6 tests this idea using A novel method with the 'blicket detector' paradigm (Gopnik & Sobel, 2000). Predictive questions ask the children to judge whether a given block will activate the detector. Diagnostic reasoning is assessed by obscuring the detector so that children cannot see which block is on it, activating it and then asking the children which block had activated it. Differences in predictive and diagnostic reasoning ability and the presence of developmental differences between 3 and 4-year-olds provides evidence in support of the hypothesis.

Finally, in Chapter 7 I summarize the key results and speculate on principles of a mechanism that can account for the variety of evidence differentiating prediction from diagnosis.

2. Asymmetries in Predictive and Diagnostic Reasoning

2.1 Introduction

We often make inferences from causes to effects and from effects to causes. A doctor might be asked for a prognosis, such as an estimate of the likelihood of some outcome given the presence of a disease, or for a diagnosis, such as an estimate of the likelihood of a particular disease given a symptom. Presumably, inferences in both directions should draw on the same knowledge; knowledge that reflects the causal structure of the events one is reasoning about as well as the strength of the causal relations. In this chapter, I use a probabilistic representation of causality to analyze how both types of judgments should change as a function of these underlying beliefs and compare people's judgments to this standard.

The focus of the chapter is on inferences that involve transmission: Some properties are likely to be transmitted from members of one category to another by virtue of a causal mechanism. Transmission requires a source (or cause) and a recipient (or effect), and inference can either go from cause to effect (the predictive direction) or from effect to cause (the diagnostic direction). To illustrate, consider reasoning about the transmission of a drug-addiction between a mother and her baby. (a) and (b) give examples of predictive and diagnostic inferences about such a scenario.

(a) A mother has a drug-addiction. How likely is it that her newborn baby has a drug-addiction?

(b) A newborn baby has a drug-addiction. How likely is it that the baby's mother has a drug-addiction?

Throughout the chapter I formalize both kinds of inferences as conditional probabilities. A predictive judgment, which I refer to as *P* is intended to be an estimate of *P(Effect | Cause)* while a diagnostic judgment, *D*, estimates *P(Cause | Effect)*.

Previous efforts to characterize inductive reasoning about transmissions have proposed that reasoning from causes to effects is more natural than from effects to causes and, on this basis, predicted that *P* should be more positively biased than *D*, all else being equal. Tversky and Kahneman (1982) report that participants rated the likelihood that a daughter has blue eyes given that her mother does to be higher than the likelihood that a mother has blue eyes given that her daughter does. They argue that the probabilities are in fact equal because the base rate probability of blue eyes should be equal across generations and therefore the conditional probabilities should also be. Medin et al. (2003) predict a *causal asymmetry* in judgments due to the ease of reasoning from causes to effects. For instance, the likelihood of lions having a property given that gazelles have it is higher than the likelihood of gazelles having it given that lions do because there is a relation of transmission from gazelles to lions through the food chain. Unlike Tversky and Kahneman, they do not analyze the normative force of their claim.

2.1.1 Determinants of Predictive and Diagnostic Likelihood

Assessing the validity of these theories requires an analysis of the conditions under which predictive judgments should be higher than diagnostic judgments because an asymmetry may arise from differences in the informational value of causes and effects rather than from psychological factors like ease or naturalness of reasoning. The drug-addiction example above exposes the issue because it violates the typical pattern. Most participants rate the diagnostic judgment to be stronger than the predictive judgment presumably

because of the differential effect of alternative causes in the two directions. The diagnostic judgment feels strong because there are few alternatives to the baby's drug-addiction besides the mother. Alternative causes should weaken *D* because they increase the likelihood that the effect was brought about by a different mechanism. They should also increase *P* for the same reason.

Another important determinant of predictive and diagnostic judgments is the probability that the cause is effective in bringing about the effect when it is present, what Cheng (1997) calls "causal power." A strong cause is more likely to bring about the effect and hence should yield higher predictive judgments. For the same reason it should also yield higher diagnostic judgments. A third factor is the prior probability of the cause in question, which should affect only diagnostic judgments. For instance, rare causes should yield low diagnostic judgments, all else being equal, because they are unlikely to have occurred. Predictive judgments should be independent of the prior probability of the cause because they should reflect only cases where the cause is present.

A normative analysis of predictive and diagnostic reasoning requires a precise formulation of the computational problem. In this chapter I provide such a formulation and model inferential strength as a joint function of all three factors. I then report the results of experiments inspired by the normative analysis.

2.1.2 Neglect of Alternatives

Though the normative analysis speaks to all three factors, the chapter's primary focus is on the effect of strength of alternative causes. A substantial literature shows that people tend to neglect alternative hypotheses when reasoning and making judgments. Using an inductive inference task with uncertain premises, Hadjichristidis, Sloman, and Over

(2009) found that people update their belief that a conclusion category has some property in a way that vastly overweights the possibility that the premise is true relative to the possibility that it is false. The effect is reminiscent of pseudodiagnosticity (Doherty et al., 1979, 1996). To test a hypothesis, people tend to choose conditional probabilities involving hypotheses that they believe true rather than conditional probabilities that would actually be diagnostic, those concerning alternative hypotheses. Using a different inductive inference task that involved making predictions about events in stories, Ross and Murphy (1996) found that participants only considered the most likely character picked out by the event, neglecting other characters. People also tend to neglect alternative causes for system failures when troubleshooting (Fischhoff, Slovic, & Lichtenstein, 1978). Over et al. (2007) show a tendency to rely on only a single possibility in causal judgments.

Reviewing the literature on how people test hypotheses and prior work on "confirmation bias" (e.g., Lord, Ross & Lepper, 1979), Klayman and Ha (1987) propose that people apply a "positive test heuristic" according to which they "test a hypothesis by examining instances in which the property or event is expected to occur (to see if it does occur), or by examining instances in which it is known to have occurred (to see if the hypothesized conditions prevail)." Evans, Over, and Handley (2003) introduced the *singularity principle* to describe this propensity to neglect alternative hypotheses. The principle implies that people will tend to focus on only a single source when making an inference.

Based on these effects and principles I expected people's inductive inferences to deviate systematically from the normative prescription. First, if people neglect alternative

causes then judgments in the predictive direction will tend to be too low and judgments in the diagnostic direction too high. Second, neglect of alternatives implies that varying the strength of alternative causes will have little effect.

2.1.3 Evidence for Considering Alternatives

In contrast to the literature showing neglect of alternatives, several studies indicate that people sometimes do take alternatives into account when reasoning diagnostically. Dougherty, Gettys and Thomas (1997) gave people vignettes describing a set of events and an outcome and asked for diagnostic judgments of the likelihood of some cause. In one example, participants read a story describing a fireman's death and judged the probability of smoke inhalation. People who thought of alternative causes for death gave lower diagnostic judgments than those who didn't, though in line with the findings above, most people tended to think of very few alternative causes.

Cummins (1995; Cummins et al. 1991) found that participants gave higher acceptability ratings to Affirming the Consequent (AC) arguments about causal scenarios with few alternative causes. For instance, an argument like, "If the trigger was pulled then the gun fired. The gun fired. Therefore the trigger was pulled." obtained high ratings relative to "If Mary jumped in the pool then she got wet. Mary got wet. Therefore Mary jumped in the pool." AC is a logical fallacy because, on the assumption "if…then" refers to a material conditional, the presence of the consequent does not imply the antecedent. Yet when interpreted causally, AC is similar to *D,* in that it requires reasoning from effect to cause. Thus these studies provide some evidence that diagnostic reasoning is sensitive to the strength of alternative causes. However, judgments of Modus Ponens, which are analogous to *P*, were insensitive to alternative strength (also see Chapter 5).

Another example comes from Waldmann (2000) who explored diagnostic reasoning in the context of a causal learning paradigm. Participants who learned about two possible diseases that could cause a symptom gave lower diagnostic judgments than those who learned about only a single cause (also see Waldmann & Holyoak, 1992).

2.1.4 Overview of the Chapter

I first propose a normative analysis of predictive and diagnostic reasoning based on Causal Bayesian networks, representations of causal structure consistent with probability theory. The analysis allows me to explore precisely how predictive and diagnostic judgments should change as a function of the causal model underlying the argument. The causal model is comprised of both a causal structure and associated parameters. A complete model specifies the strength of alternative causes, prior probabilities, and causal power. Experiments 1 and 2 test how people's judgments compare to the normative analysis by collecting predictive and diagnostic judgments along with the primitives for those judgments, the model parameters. Experiments 3 and 4 address an alternative pragmatic explanation for the results of Experiments 1 and 2. In the general discussion I compare the analysis to other models of property induction and discuss the implications of the findings.

2.2. Normative Causal Model Analysis

The goal of the normative analysis is to capture the contribution of alternative causes, causal power, and prior probability to predictive and diagnostic reasoning in a way that is probabilistically coherent. I accomplish these requirements using the Causal Bayes net framework (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). Causal Bayes nets are graphical representations of causal structure that are defined in terms of probabilistic

independence and interventional logic (Sloman, 2005; Woodward, 2004). In the graph, nodes are variables that represent events or properties and edges represent the causal relations among the nodes. Causal Bayes nets can be used to compute the probabilities of unobserved nodes given observation of or interventions on other nodes and can therefore concisely represent the desired conditional probability distributions, *P* and *D*.

A transmission argument can be represented by a common-effect structure, one effect with multiple possible causes. In general, a predicate might be transmitted to the effect category from the target cause, or by some alternative cause. To capture the additional constraint that a true alternative cause should be independent of the target cause I restrict myself to arguments in which transmission from a source to a recipient follows an independent causal path. Kelley (1972) proposed the "multiple sufficient causes" schema to describe independent causes that combine to generate an effect according to an inclusive-or function. Any of the causes is individually sufficient to bring about the effect, and if more than one cause is present, the effect is also present. The probabilistic extension of the inclusive-or is called the noisy-or function. The presence of either cause raises the probability of the effect and if both causes are present the probability of the effect is even higher, increasing according to the independent contribution of each cause. When the noisy-or model applies, the calculations of *P* and *D* specified by the model are the only ones that are consistent with the parameters. In that sense the model offers a normative benchmark for arguments that concern an appropriate causal model. I chose arguments to satisfy the necessary conditions: target and alternative causes were each sufficient for the effect (though only effective some of the time) and as independent from each other as possible.

2.2.1 Model Description

A Causal Bayes net can be fully described by the probability distributions of its exogenous variables, variables that have no parents in the graph, along with a set of functions and parameters that define the probability distributions of endogenous nodes conditioned on their parents. In other words, the model requires specifying the prior probability distributions of all root causes and functions describing how causes combine to generate effects.

By aggregating all alternative causes into a single node, a causal background (Cheng, 1997), the structure necessary for defining *P* and *D* can be concisely represented as a causal Bayes net with three nodes: the cause, the effect and the aggregate of all alternative causes. Separate edges connect the cause and alternative to the effect. To specify the parameters over this structure I assumed that events are binary; they either happen or they do not. This allowed me to represent the probability distribution of exogenous nodes with a single number, a prior probability. I also assumed that the cause and any alternative causes are independent and generate the effect independently according to a nosiy-or function as discussed above. The independent contribution of a cause can be defined in the model as a parameter that specifies the conditional probability of the effect given that cause and no others (a 'causal power'). Because of its use of the noisy-or function and parameterization in terms of causal powers, the structure is identical to that proposed in Cheng's seminal PowerPC model of causal learning.[2]

To simplify calculations, I collapsed the prior probability and causal power of the alternative causes into a single parameter denoting the strength of alternatives, set to $P(Effect \mid \sim Cause)$. This is akin to setting the prior to one (i.e. assuming alternatives are

always present but only effective in bringing about the effect some of the time.) The prior and causal power of alternatives are always confounded in the model, so the simplification is not substantive.

The model is therefore fully parameterized by three numbers: the prior probability of the cause (P_c), the causal power of the cause (W_c) equal to *P(Effect | Cause, ~Effective Alternative Causes),* and the strength of alternatives (W_a) or *P(Effect | ~Cause)*. The structure and parameterization are depicted in Figure 2.1. In the figure W_a represents both the prior and causal power of alternatives collapsed into a single term.

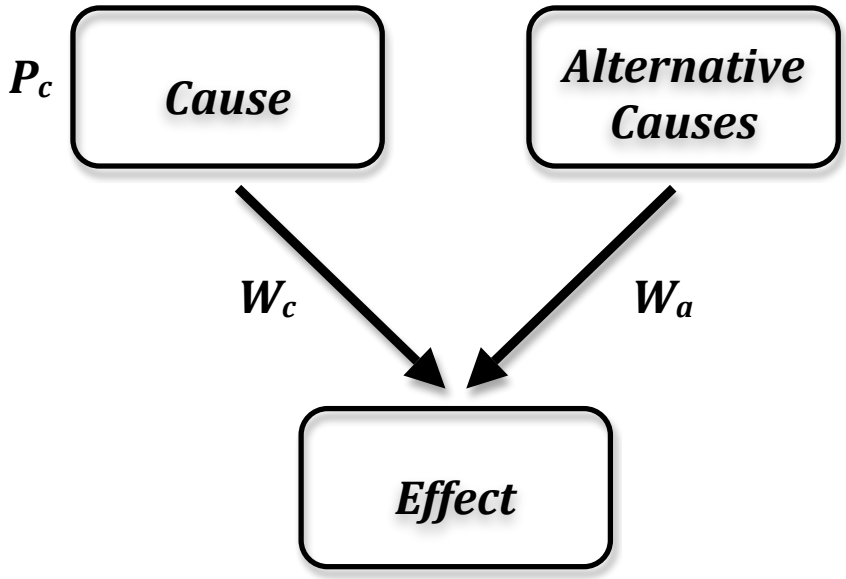

Figure 2.1: A Bayes net model of transmission arguments. P_c represents the prior probability of the cause, W_c is the causal power of the cause and W_a is the strength of alternatives, the aggregate causal power and prior probabilities of all alternative causes collapsed into a single term. The effect is generated by a noisy-or function of the cause and the alternatives.

The predictive judgment (*P)* and diagnostic judgment (*D*) correspond to *P(Effect | Cause)* and *P(Cause | Effect)*, respectively. *P* is calculated by direct application of the noisy-or equation:

$$P = P(Effect \mid Cause) = W_c + W_a - W_c W_a \quad (1)$$

Note the difference between W_c and P. P represents the probability that the effect occurs given that the cause occurred. This includes both cases in which the cause was effective in generating the effect and cases in which the cause was not effective but an alternative cause was. Therefore, P is higher than W_c and increases with the strength of alternatives.

The diagnostic judgment, D, is derived by considering its complement, the probability that the cause *did not* occur despite the effect having occurred (for an alternative derivation see Waldmann et al. 2008).

$$D = P(Cause \mid Effect) = 1 - P(\sim Cause \mid Effect) \tag{2}$$

By Bayes' rule:

$$D = 1 - P(Effect \mid \sim Cause)\frac{P(\sim Cause)}{P(Effect)} = 1 - p(\sim Cause)\frac{P(Effect \mid \sim Cause)}{P(Effect)} \tag{3}$$

Deriving *P(Effect)* by the noisy-or equation and substituting W_a for *P(Effect | ~Cause)* and (1 - P_c) for *P(~Cause)*:

$$D = 1 - (1 - P_c)\frac{W_a}{P_c W_c + W_a - P_c W_c W_a} \tag{4}$$

Equation 4 shows that two factors determine D, the prior probability of the cause and the probability that the alternatives caused the effect (i.e. the ratio between W_a and the extension of *P(Effect)* at the end of Equation 4). The presence of the effect cannot decrease the probability of the cause, so D is always higher than P_c and it increases with P_c Conversely, the effect is diagnostic of the cause to the extent it was not generated by alternative causes. Therefore, the cause and the alternatives compete to explain the effect and D decreases with the probability that the alternative causes caused the effect.

2.2.2 Model Predictions

Equations 1 and 4 yield predictions regarding how judgments of P and D should vary as a function of the parameters P_c, W_c and W_a. P is a function of two parameters, W_c and W_a, and increases as each of them increases independently. D is a more complex function of all three parameters; it depends on the prior probability of the cause and the probability that the effect was caused by the alternatives. The probability that the effect was caused by the alternatives is a comparative measure of the strength of alternatives relative to the strength of the focal cause. Accordingly, it increases with W_a and decreases with P_c and W_c. Therefore, D increases as P_c or W_c increases or as W_a decreases.

2.3 Experiment 1

In Experiment 1 I compared predictive and diagnostic judgments about arguments in which there are either strong or weak alternative causes, and manipulated the strength of alternatives by keeping the categories constant while varying the predicate. Alternative causes, prior probability and causal power were never mentioned explicitly so the experiment tested people's ability to use aspects of their intuitive causal models in generating likelihood judgments. According to the normative analysis, all else being equal, P should increase with strong alternatives while D should decrease. If people neglect alternative causes then varying the strength of alternatives should have little effect on P or D.

The ultimate goal of Experiment 1 was to generate enough data to test whether the normative model accounts for people's predictive and diagnostic judgments. I therefore collected judgments of the model parameters P_c, W_c and W_a along with predictive and diagnostic judgments. If people's inductive judgments are consistent with their beliefs

about the relevant probabilities then the conditional probabilities derived from the parameters according to the model should match the predictive and diagnostic judgments.

I relied on pre-existing causal beliefs rather than train people on novel causal systems (e.g. Rehder, 2006). Collecting all of the parameters and fitting the model alleviates some of the concerns associated with using naturalistic materials. Ideally, the items would not vary systematically in the other parameters across the manipulation, and I used a large number of items to try to make that likely. Nonetheless, using naturalistic materials, potential confounding is always a concern. The model fitting allowed us to interpret the results even in the case of confounding. Thus the effects across conditions are only suggestive. It is the modeling that provides the real interpretive power.

Another concern with using people's pre-existing beliefs is that one cannot be certain how well those beliefs conform to the model assumptions. The primary concern is that the main cause and the alternative causes might not be completely independent. This is a valid concern, but is mitigated by the fact that alternative causes are necessarily probability raising. Therefore, even if dependence is introduced, the normative value for P is still higher than W_c, unless the causes are perfectly correlated. I chose materials with the independence assumption in mind, so on average the value for P should be close to the model prediction. An analogous argument applies to judgments of D.

2.3.1 Methods

Participants

162 participants were recruited by Internet advertisement and participated online for a chance to win a $100 lottery prize. Additionally, 18 Brown University students participated in the lab for class credit or were paid at a rate of eight dollars per hour. In

total 180 participants completed the experiment. Internet participants were recruited on college message boards and logged on to the survey remotely. Lab participants were recruited through the Brown University psychology research pool or through flyers posted on campus and completed the questionnaire on a computer in the lab.

Design

The experiment had three independent variables: categories, strong versus weak alternatives and question type. Categories and predicates were chosen to fit the common effect noisy-or causal structure where any alternative causes provide an independent contribution to the effect and the causal relation from cause to effect is unidirectional. For each predicate I asked five questions: the prior probability of the cause (P_c), the causal power of the cause (W_c), the strength of alternatives (W_a), the predictive judgment (P) and the diagnostic judgment (D). To probe these I simply asked for the likelihood of the relevant events on a 0-100 scale. Examples of the question forms are shown in Table 2.1. I chose 20 sets of categories, two predicates for each set, and five questions for each predicate for a total of 200 questions. The predicates and categories are shown in Appendix A.[3]

Table 2.1: *Example Question Forms from Experiment 1*

Parameter / Judgment	*Example Wording*
Prior Probability of Cause (P_c)	A woman is the mother of a newborn baby. How likely is it that the woman is drug-addicted?
Causal Power of Cause (W_c)	The mother of a newborn baby is drug-addicted. How likely is it that her being drug-addicted causes her baby to be drug addicted?
Strength of Alternatives (W_a)	The mother of a newborn baby is not drug addicted. How likely is it that her baby is drug addicted?
Predictive Judgment (*P*)	The mother of a newborn baby is drug-addicted. How likely is it that her baby is drug-addicted?
Diagnostic Judgment (*D*)	A newborn baby is drug addicted. How likely is it that its mother is drug addicted?

To avoid interactions among questions about the same predicate, the 200 questions were each assigned to one of five questionnaires of 40 questions each. Each participant received one questionnaire. Questions were randomly assigned with the constraints that each questionnaire had one question type from each of the 40 predicates and that no questionnaire had the same question type of the weak and strong predicate for a given set of categories. Each participant therefore answered a single question about each predicate. The order of questions in each questionnaire was randomized but constant for each questionnaire.

Materials and Procedure

Participants were randomly assigned to one of the five questionnaires. Each questionnaire consisted of instructions at the top followed by 40 questions, all on a single sheet. Participants were instructed to "Give an answer between 0 (impossible) and 100 (definite)" for each question. The experiment took approximately 20 minutes.

2.3.2 Results

Five participants gave the same response to all 40 questions and were omitted from subsequent analysis. The mean predictive and diagnostic judgments for the strong and weak alternatives conditions are shown in Figure 2.2. I collapsed the data across participants and assessed the relative effect of strength of alternatives on predictive and diagnostic judgments by performing a 2 (alternatives: strong vs. weak) x 2 (judgment: predictive vs. diagnostic) repeated measures ANOVA.[4] There was a significant interaction between judgment type and strength of alternatives, $F(1,19)=31.4$, $p<0.00001$, partial $\eta^2=0.62$. There was also a main effect of strength of alternatives, $F(1,19)=4.9$, $p=0.039$, partial $\eta^2=0.21$, but no significant effect of type of judgment, $F(1,19)=0.6$, *ns*.

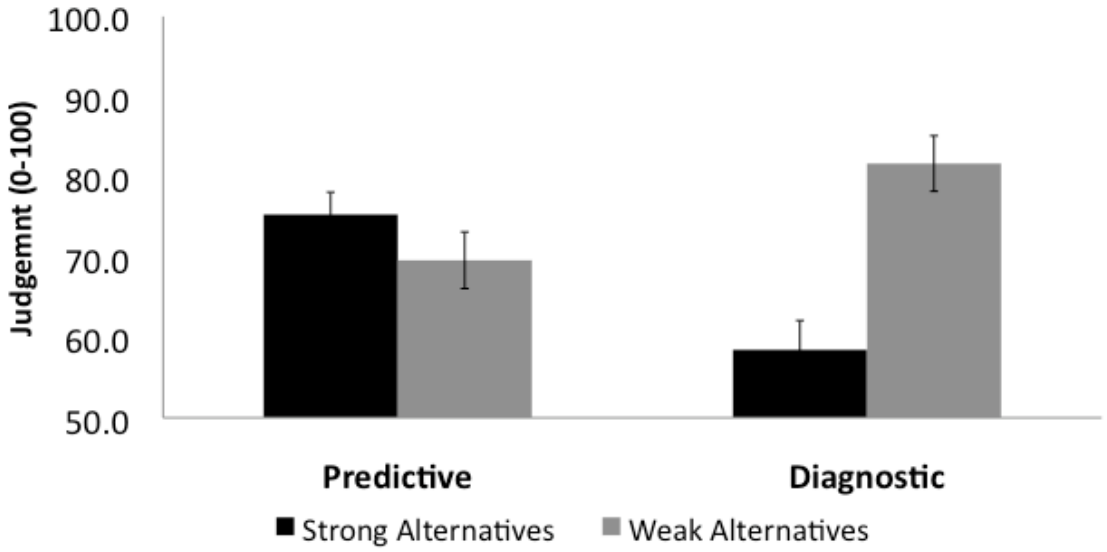

Figure 2.2: Mean Parameter Judgments for the Strong and Weak Alternatives Conditions of Experiment 1

I conducted planned comparisons between judgments in the strong and weak alternatives conditions. Diagnostic judgments in the weak alternatives condition (M = 81.7) were higher than in the strong alternatives condition, M= 58.5; $t(19) = 5.0$, $p<0.00001$, Cohen's d=1.1. Predictive judgments did not differ significantly, $M_{strong}=75.3$; $M_{weak}=$

69.6; $t(19) = 1.31$, *ns*. I also used matched sample t-tests to compare mean parameter judgments for each category set across the strong/weak manipulation. The results are shown in Table 2.2. The manipulation of strong vs. weak alternatives was effective as evidenced by the difference between W_a in the two conditions. P_c and W_c[5] responses didn't differ significantly between conditions.

Table 2.2: Mean Parameter Judgments for the Strong and Weak Alternatives Conditions of Experiment 1

Parameter	*Strong Alternatives*	*Weak Alternatives*	*t-stat*	*p-value*
Prior Probability of Cause (P_c)	41.6	48.2	1.14	0.3
Causal Power of Cause (W_c)	75.0	71.4	0.79	0.4
Strength of Alternatives (W_a)	39.0	20.0	5.00	<0.00001

2.3.3 Model Fits

Modeling Details

The model represents the relation between a single participant's judgments of the parameters P_c, W_c and W_a and their judgments of P and D. Because of the incomplete design, no participant made all of the parameter judgments for any single item, and I therefore had a distribution of unmatched judgments of the parameters for each item. I could not simply take the means of these distributions and combine them according to the model's equations because it is not generally true that the mean of a function of distributions is equivalent to applying that function to their means. In particular, the equation for D, which includes random variables in the denominator, violates this assumption. For P the assumption did hold, and the model's outputs for P were the same as if they were calculated directly from the parameter means. Nonetheless, for consistency's sake I used the same procedure to generate predictions for P and D.

My method was to use a sampling procedure to generate a distribution for the model's predictions of *P* and *D* for each item and used the mean of this distribution as the model's prediction for that item. To generate a single sample of *P* and *D* for a given item I drew one sample of each of the three parameters uniformly and independently from the set of participant responses. I then calculated *P* and *D* from the sampled parameters according to Equations 1 and 4. I repeated this procedure to generate 100,000 samples each of *P* and *D* for each item and took the means as the model's predictions for that item. Reruns of the sampling procedure yielded no differences in the predictions for either *P* or *D*.

Modeling Results

Figure 3 shows the model predictions for *P* (left panel) and *D* compared to participant responses. As with participant responses, model predictions for *D* were higher in the weak condition (M=0.79) than in the strong condition, M=0.61; $t(19)= 5.0$, $p<0.00001$, Cohen's d=1.0. Model predictions for *P* were lower in the weak condition (M=0.77) than in the strong condition, M=0.85; $t(19)=2.38$, $p=0.028$, Cohen's d=0.5. The model predictions of *D* were not significantly different from participant responses, $t(39)=0.71$, $p=0.48$, Cohen's d=0.12 , and were highly correlated with items in the strong and weak conditions separately, $r_{strong}=0.69$ $p<0.00001$; $r_{weak}=0.69$, $p<0.00001$, and across both conditions, $r=0.80$, $p<0.00001$. Model predictions of *P* (M=0.81) were significantly higher than participant responses, M=0.72; $t(39)=6.54$, $p<0.00001$, Cohen's d=1.09, but were still highly correlated both within each condition, $r_{weak}=0.83$, $p<0.00001$; $r_{strong}=0.75$, $p<0.00001$, and across conditions, $r=0.72$, $p<0.00001$.

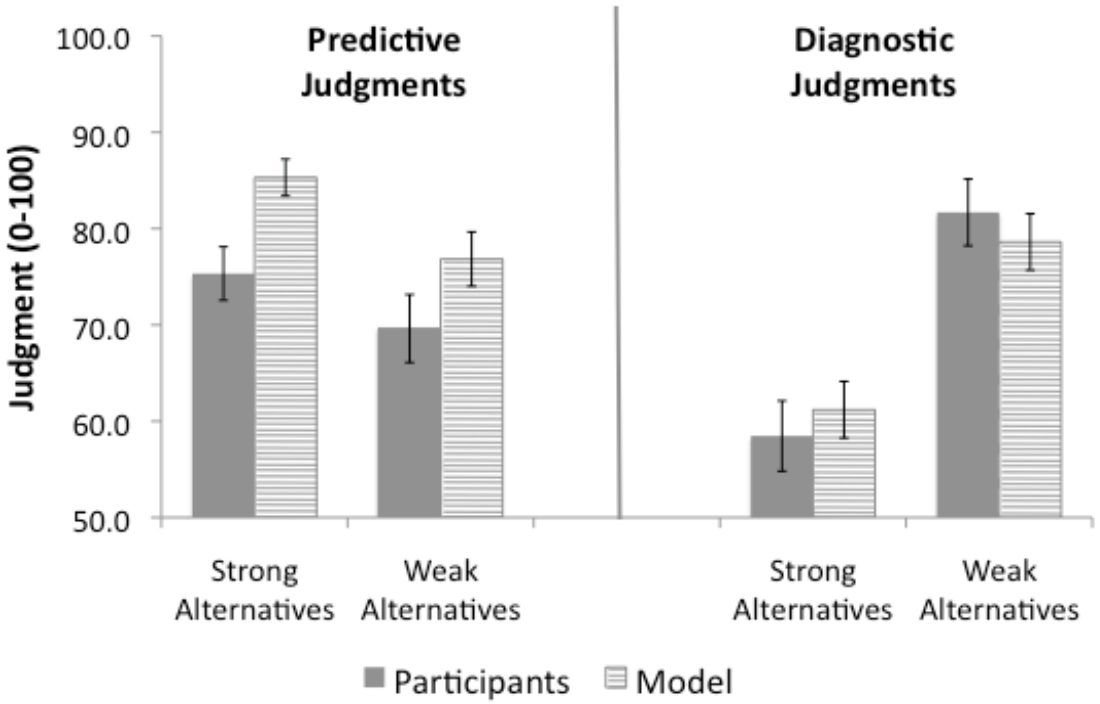

Figure 2.3: Comparisons between mean participant responses and model predictions for Experiment 1 with standard errors. Predictive judgments are shown in the left panel and diagnostic judgments on the right.

A possible concern is that the normative model is superfluous and that one of the parameters alone can predict judgments of *P* and *D*. I therefore used hierarchical multiple regression analyses to test whether the normative model does better than individual parameters at accounting for the variance in *P* and *D* judgments across items. The results of those analyses are shown in Table 2.3. For judgments of *D* I considered the possibility that the high correlation between the model and judgments of *D* could be driven primarily by differences in W_a. W_a was significantly correlated with *D* across the strong/weak manipulation, r=-0.49, p=0.003, however the correlations were not significant in each condition separately, r_{weak}=-0.28; r_{strong}=-0.08. The hierarchical multiple regression, which used W_a and the normative model as predictors of *D* showed that the model fit the data better than W_a alone and W_a had no predictive value beyond its role in the model.

Together, the normative model and W_a accounted for 64% of the variance in *D*. The unique variance of the normative model accounted for 41% of the variance of *D*, F(1,39)=41.7, $p<0.00001$, but the unique variance of W_a did not account for any of the variance of *D*, F(1,39)=1.0, *ns*.

In contrast, the best predictor of predictive judgments was the single parameter W_c and not the full model. W_c alone fit the data better than the model and the model had no predictive value beyond that of W_c. The model and W_c together accounted for 77% of the variance of *P*. The unique variance of W_c accounted for 10% of the variance of *P*, F(1,39)=17.1, $p<0.00001$, but the unique variance of the model did not account for any of the variance of *P,* F(1,39)=0.4, *ns*. Because W_c and W_a are the only two factors in the model prediction of *P* these results imply that predictive judgments were uncorrelated with W_a, which I verified, $r=0.044$, $p=0.78$. Corroborating this analysis I also found that there was no significant difference between judgments of *P* and W_c, $t(39)=0.60$, *ns*.

Table 2.3: Variance of Predictive and Diagnostic Judgments Accounted for by the Normative Model Versus a Single Predictive Parameter

Predictor	*All Variance*	*Unique Variance*	*p-value of Unique Variance*
Diagnostic Judgments			
Strength of Alternatives (W_a)	0.23	0.01	0.32
Model Prediction for D	0.63	0.41	<0.00001
Predictive Judgments			
Causal Power (W_c)	0.77	0.10	<0.00001
Model prediction for P	0.67	0.002	0.53

2.3.4 Discussion

Participants were sensitive to alternative strength when reasoning diagnostically but not predictively. I found a large difference of alternative strength for diagnostic judgments

but no difference for predictive judgments. The model fitting allowed me to rule out possible alternative explanations for this pattern. When predictive judgments were extrapolated using the model, the results were significantly underestimated by the predictive judgments that were probed directly. This underestimation was driven by the lack of consideration of W_a. Predictive judgments were invariant to W_a and were similar to W_c, judgments of causal power.

The model achieved good fits to participants' diagnostic judgments, with zero free parameters. The model did not just achieve a good fit when the data were aggregated over arguments. Instead, the model accounted for a large part of the variance across specific arguments. The model's good fit did not simply capture participants' sensitivity to alternative strength. The model was highly correlated with participant judgments within the strong and weak conditions separately while W_a was uncorrelated with those judgments and W_a had no predictive value beyond its role in the model. In other words, the strength of alternatives was only important in the context of the other parameters. On average, participants combined information about prior probability, causal power, and alternative strength in a way that approximated the normative computation fairly closely.

2.4 Experiment 2

Due to the partially between-participants design, the model fitting in Experiment 1 required generating samples from the posterior distribution of D as opposed to calculating model fits directly based on the parameters given by participants. The purpose of Experiment 2 was to replicate the findings of Experiment 1 with a design that allowed me to calculate predicted values of P and D directly from each participant's judgments. This meant that all parameter estimates had to be collected from each participant. I also

collected judgments of *P(Effect)* or P_e (e.g. A woman is the mother of a newborn baby. How likely is it that the newborn is drug-addicted?). This allowed me to compare the model to an alternative model of categorical induction.

2.4.1 Methods

Participants

78 participants were recruited by Internet advertisement and participated online for a chance to win a $100 lottery prize. Additionally, 30 Brown University students participated in the lab for class credit or were paid at a rate of eight dollars per hour. In total 108 participants completed the experiment.

Design

I chose five sets of categories from the 20 that were used in Experiment 1. As in Experiment 1, strong versus weak alternatives was also manipulated by choosing two different predicates for each set of categories and question type was a third independent variable (I collected judgments of P_c, W_c, W_a, P, D and P_e). There were five category sets, two predicates for each category and six questions for each predicate for a total of 60 questions. All variables were manipulated within participant so each participant answered all 60 questions.

To attenuate interactions among items I split the questions onto three pages so that each predicate was represented in two questions per page, W_c and D, P_c and P, or W_a and P_e. The order of questions on each page was randomized. To test for order effects, I created two versions of the questionnaire. The second version displayed the questions and pages in reverse order from the first version. Participants were randomly assigned to one of the two versions.

Procedure and Stimuli

I chose five of the category sets from Experiment 1: Mother/Baby, Apple Slices/Apple Pie, Football Coach/Team, Engine/Honda Accord, and Music/Party. The questions were the same as in Experiment 1 except, first, the wording of the diagnostic question for the weak alternatives coach/team predicate was changed to a more natural form. I also changed the strong alternatives predicate for the engine/Honda Accord question because of the concern that the engine not functioning properly implies that the car does not function properly. I therefore used the predicate 'is noisy' instead.

The procedure was identical to Experiment 1 except that there were 60 questions instead of 40 and they covered three pages rather than one. I also added the following to the instructions: "Please answer the questions in order. Once you've answered a question don't go back and change it. Though some of the questions are similar to previous questions, it is important to answer every question in the set." The questionnaire took approximately 30 minutes to complete.

2.4.2 Results

One participant gave the same response to each question and was omitted from subsequent analyses. Responses to the two question orders were highly similar, $r=0.98$, $p<0.00001$. The responses of internet and lab participants were also highly similar, $r=0.98$, $p<0.00001$. All subsequent analyses therefore use the full data set collapsed over orders and internet/lab populations.

The mean predictive and diagnostic judgments for the strong and weak alternatives conditions are shown in Figure 4. I subjected the participant means to a 2 (alternatives: strong vs. weak) x 2 (judgment: predictive vs. diagnostic) repeated measure

ANOVA. Once again I observed a significant interaction between alternative strength and judgment type, $F(1,106)=137.7$, $p<0.0001$, partial $\eta^2=0.57$. There was also a main effect of alternative strength, $F(1,106)=6.3$, $p=0.014$, partial $\eta^2=0.056$, but no main effect of question type, $F(1,106)=0.72$, *ns*.

Planned comparisons between judgments in the strong and weak alternatives conditions revealed that diagnostic judgments in the weak alternatives condition (M = 83.5) were higher than in the strong alternatives condition, M= 70.4; $t(106) = 9.13$, $p=<0.00001$, Cohen's d=0.88. Unlike Experiment 1, predictive judgments were significantly higher for strong items than weak ones, $M_{strong} = 80.2$; $M_{weak} = 72.2$; $t(106) = 6.3$, $p<0.00001$, Cohen's d=0.63.

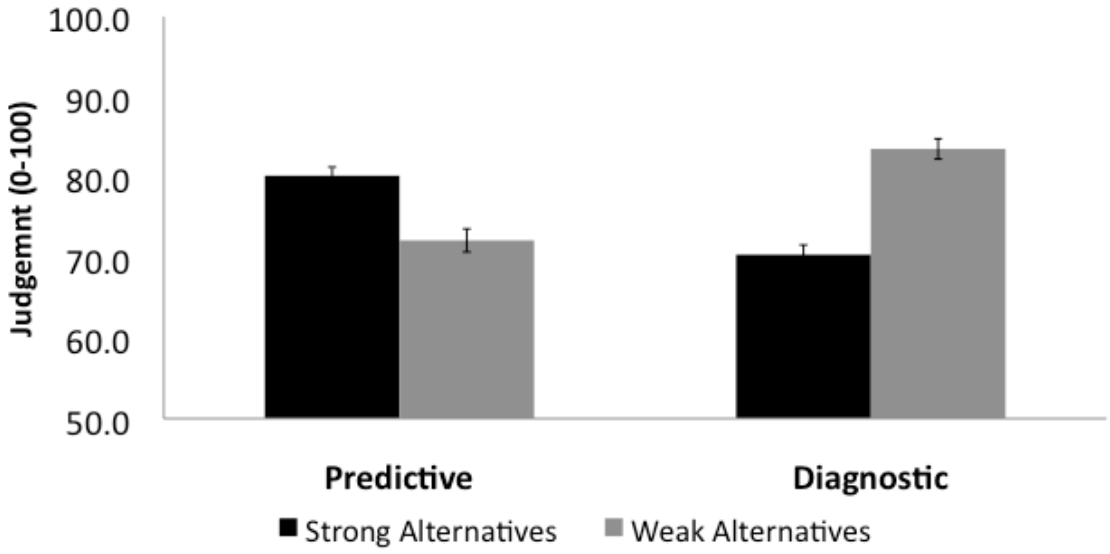

Figure 2.4: Mean Predictive and Diagnostic judgments and standard errors for the strong and weak alternatives conditions of Experiment 2.

To assess any parameter differences across the strong/weak manipulation, I performed matched-sample t-tests on question means (Table 2.4). Replicating

Experiment 2, W_a was judged higher for the strong items than the weak ones. P_c, W_c and P_e were also judged higher for strong items than weak ones. Due to the parameter differences between conditions, no conclusions about the relative neglect of alternatives for predictive and diagnostic judgments can be drawn without model fitting.

Table 2.4: Mean Parameter Judgments for the Strong and Weak Alternatives Conditions of Experiment 2

Parameter	*Strong Alternatives*	*Weak Alternatives*	*t-stat*	*p-value*
Prior Probability of Cause (P_c)	50.6	42.4	7.57	<0.00001
Causal Power of Cause (W_c)	78.5	74.7	2.90	0.005
Strength of Alternatives (W_a)	38.9	17.2	15.62	<0.00001
Prior Probability of Effect (P_e)	49.4	34.6	11.45	<0.00001

2.4.3 Model Fits

Because of the within-participants design I was able to calculate model predictions for each participant and each item instead of sampling as I did in the analysis of Experiment 1. For each participant I simply took the parameters they gave for a particular item and calculated Equations 1 and 4 to yield a prediction for *P* and *D* for that item.

Figure 2.5 shows model predictions compared with participant responses. As in Experiment 1, the model overestimated participants' predictive judgments in both the strong, $t(106)=9.6$, $p<0.00001$, Cohen's d=1.0 and weak conditions, $t(106)=6.4$, $p<0.00001$, Cohen's d=0.4. The model fits for diagnostic inferences were much closer. In the strong condition model predictions and participant judgments were not significantly different, $t(106)=1.3$, $p=0.19$, *ns*. In the weak condition participant responses were lower than the model, but this difference was very small, $t(106)=2.1$, $p=0.04$, Cohen's d=0.2.

As in experiment I conducted a hierarchical multiple regression analysis, this time using participant responses collapsed over categories. The analysis revealed the same pattern as in Experiment 1. The variance of diagnostic judgments was better accounted for by the model than by W_a with the unique variance of the model accounting for 12% of the variance in D, p<0.00001, but the unique the unique variance of W_a not accounting for any variance in D, p=0.49. The variance of predictive judgments was better accounted for by W_c than by the model. The unique variance of Wc accounted for 6% of the variance in P, p<0.00001, but the unique variance of the model did not account for any variance in P, p=0.39. Once again, W_a was uncorrelated with predictive judgments (r=0.03, p=0.85).

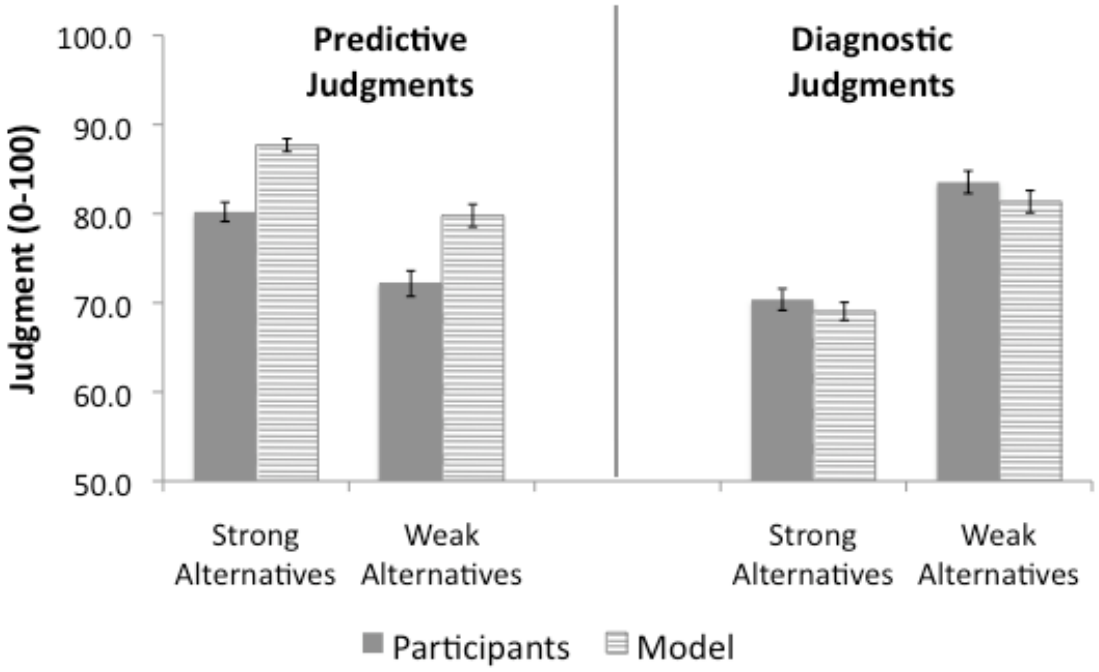

Figure 2.5: Comparisons between mean participant responses and model predictions for Experiment 2 with standard errors. Predictive judgments are shown in the left panel and diagnostic judgments on the right.

2.4.4 Discussion

The results of Experiment 2 corroborated the conclusions of Experiment 1. The pattern of results was somewhat different than Experiment 1 as predictive judgments were significantly higher in the strong alternatives condition than the weak alternatives condition. However, the model fitting showed that this difference was not due to differences in W_a. Once again, predictive judgments were uncorrelated with alternative strength and were lower than the predictions of the model. The differential pattern from Experiment 1 was likely due to the small number of categories used in the experiment. Also corroborating Experiment 1, the model predicted diagnostic judgments more closely.

2.5 Experiment 3

The conclusion from Experiments 1 and 2 that participants neglected alternatives in the predictive direction is based in part on the similarity between predictive judgments and judgments of causal power, W_c. I attribute this to how people reason but it could instead reflect how they interpreted the probe questions. One possibility is that participants may have interpreted the W_c question as asking for P. The W_c question asks participants to judge the likelihood that the cause causes the effect. Participants might not understand this question as asking for causal power and give a conditional probability judgment instead.

In Experiment 3 I tested this possibility by mentioning an alternative cause explicitly and then asking the W_c and P questions. I expected participants to take the mentioned alternative into account and give higher P judgments than W_c judgments as per

the normative model. The misinterpretation hypothesis predicts that judgments of P and W_c should be the same even when alternatives were mentioned.

2.5.1 Methods

Participants

62 Brown University students were recruited on campus and participated voluntarily. 31 were assigned to each condition.

Design

I chose ten of the strong alternative items from Experiment 1 to maximize the effect of mentioning the alternative cause. The main independent variable was whether participants were asked for judgments of P or W_c and it was manipulated between participants. Each participant therefore answered either ten P questions or ten W_c questions. All of the questions explicitly mentioned the possibility of an alternative cause without saying whether that cause was present. An example of a P question is "The coach of a high school football team is highly motivated. Accolades from family and friends could also cause high school football teams to be highly motivated. How likely is it that the team is highly motivated?" The analogous W_c question was "The coach of a high school football team is highly motivated. Accolades from family and friends could also cause high school football teams to be highly motivated. How likely is it that the coach being highly motivated causes his team to be highly motivated?"

Procedure and Stimuli

Participants were handed a single sheet with the ten questions and instructions at the top. The questionnaire took between five and ten minutes to complete. The stimuli used in the experiment are shown in Appendix A.

2.5.2. Results

Due to a typographical error in one of the questionnaires, the "Honda Accord" item was omitted from the analysis. The mean *P* and W_c judgments for Experiment 3 are shown in Figure 2.6 along with those for the same items from Experiment 1 for comparison. An independent sample t-test on participant means revealed that judgments of *P* (M=81.7) were significantly higher than W_c, M=72.0, $t(60)=3.49$, $p=0.0009$, Cohen's d=0.9, as predicted by the neglect hypothesis. A matched sample t-test on category means yielded the same result.

I also compared the results to those for the same items from Experiment 1. An independent sample t-test on participant means showed that predictive judgments were significantly greater in Experiment 3 than Experiment 1, $\text{Mean}_{\text{exp1}}=71.9$, $t(204)=3.56$, $p=0.0005$, Cohen's d=0.5, but that judgments of W_c were not different across experiments, $\text{Mean}_{\text{exp1}}=73.9$, $t(204)=0.61$, *ns*. The pattern was the same when collapsed over participants.

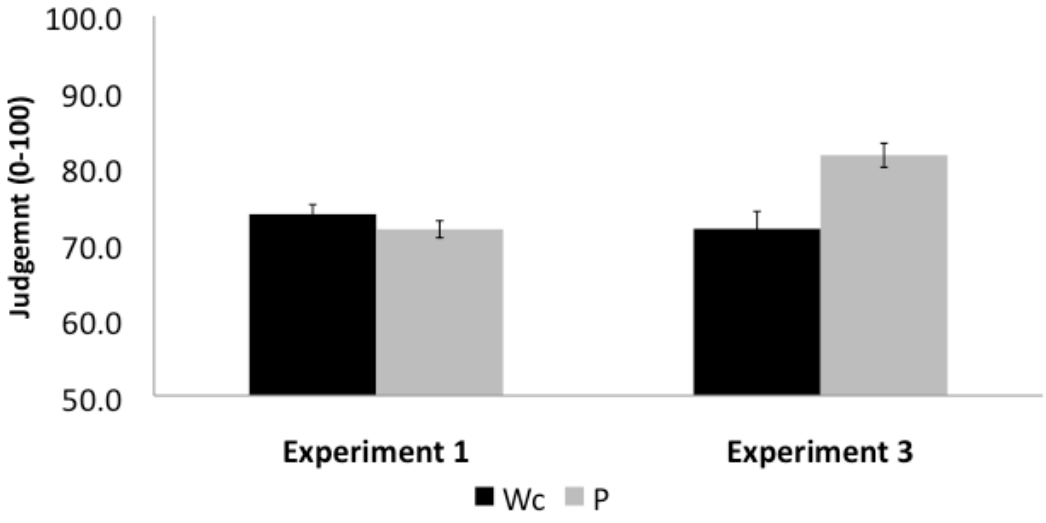

Figure 2.6: Mean P and W_c judgments for Experiment 3, with the judgments for the same items from Experiment 1.

2.5.3 Discussion

In Experiment 3, judgments of *P* were higher than W_c when the possibility of an alternative cause was mentioned explicitly. Judgments of W_c were similar to judgments of both *P* and W_c for the same items from Experiment 1. This suggests that participants took alternative causes into account in judging *P*, but only when alternatives were mentioned explicitly. The increase in judgments of *P* was not brought about by giving people new information, but rather by directing their attention to something they already knew. For example, most participants were likely aware that accolades from family and friends might motivate high school football teams.

These results rule out the possibility that participants are answering the *P* question when they are asked the W_c question. However, there is an additional possibility that participants interpret the *P* question as asking for W_c. Experiment 3 speaks against this possibility because participants treated the questions differently, but doesn't rule it out. The possibility remains that participants understand that they should take into account not only causes mentioned in the question itself, but also those mentioned in the context of the question, even if those causes aren't definitively present. Thus in Experiment 3, they might have interpreted the *P* questions as asking for the probability of the effect conditioned on the presence of the main cause and the possibility of the alternative cause mentioned, but no other causes, which would have led to higher judgments than for the W_c questions. Extending the pragmatic hypothesis in such a way makes it much more difficult to pin down and differentiate from neglect of alternatives. Experiment 4 took a different approach to addressing it.

2.6 Experiment 4

In Experiment 4 I address the pragmatic account differently, by reducing the vagueness of the questions. Hertwig and Gigerenzer (1999) have shown that people have a variety of different interpretations of probability and that questions about frequency are less vague. Therefore, instead of asking participants for the likelihood of the effect given the cause in this experiment, I specified a definite set of instances and asked participants to estimate the frequency of a subset. Consider the following example:

> (a) Consider mothers who each have a single newborn baby. Of 100 mothers who are drug-addicted, how many of the mothers' babies are drug-addicted?

The question asks for the number of babies out of 100 that are drug-addicted. To interpret this question as asking for W_c would imply that one should not include drug-addicted babies whose drug-addiction is due to some other source besides the mother. This would be an odd interpretation given that the question explicitly asks for the number of drug-addicted babies. A further benefit of frequency formats is that, under some conditions, they are one way to obtain more veridical representations of uncertainty (Barbey & Sloman, 2008; Gigerenzer & Hoffrage, 1995). Experiment 4 thus serves to test the robustness of the inductive asymmetry.

2.6.1 Methods

Participants

68 undergraduates from the Brown University psychology pool participated for class credit.

Design, stimuli and procedure

I asked three types of questions: *P, D* and W_c. *P* questions were phrased as in the example in (a). *D* and W_c questions were phrased as in (b) and (c) respectively:

(b) Consider mothers who each have a single newborn baby. Of 100 babies who are drug addicted, how many of the babies' mothers are drug-addicted?

(c) Consider mothers who each have a single newborn baby. Of 100 mothers who are drug-addicted, in how many cases does the mother being drug-addicted cause her baby to be drug-addicted?

I utilized all 20 category sets and the strong and weak predicates from Experiment 1. The 120 questions were divided into three questionnaires such that no questionnaire had the strong and weak version for a particular question type. Each participant was assigned at random to one of the three questionnaires and completed the experiment in approximately 20 minutes.

2.6.2 Results

The results of Experiment 4 are depicted in Figure 2.7. I collapsed over categories and subjected the data to a 2 (predictive/diagnostic) X 2 (strong/weak) ANOVA. Replicating Experiment 1 there was a significant interaction between strength of alternatives and direction of inference, $F(1,67)=46.0$, $p<0.00001$, partial $\eta^2=0.4$. There was also a main effect of direction of inference, $F(1.67)=9.5$, $p=0.003$, partial $\eta^2=0.4$, and strength of alternatives, $F(1,67)=43.4$, $p<0.00001$, partial $\eta^2=0.1$. Collapsing the data over participants and comparing question means yielded a similar pattern: a significant interaction, $F(1,19)=19.1$, $p<0.00001$, partial $\eta^2=0.5$ and a main effect of direction of inference, $F(1,19)=5.6$, $p=0.029$, partial $\eta^2=0.2$ but no main effect of strength of alternatives, $F(1,19)=2.0$, *ns*.

I performed a series of planned comparisons to test the impact of strength of alternatives. Replicating Experiment 1, there was a large difference between judgments of *D* across the strong/weak manipulation, M_{strong}=69.0 M_{weak}=86.8; t(67)=8.1, p<0.00001, Cohen's d=1.0, but no difference for judgments of *P*, M_{strong}=82.2 M_{weak}=81.5; t(67)=0.5, *ns*. Judgments of *P* and W_c did not differ for either the strong (M_{wc} =79.5; t(67)=1.1, *ns*) or weak (M_{wc} =78.8; t(67)=1.3, *ns*) items. Collapsing the data over participants and comparing question means yielded the same results: a large difference between judgments of *D* across the strong/weak manipulation, t(19)=4.9, p=0.0001, Cohen's d=1.1, no difference for judgments of *P,* t(19)=0.6, *ns*, and no difference between *P* and W_c for either strong (t(19)=1.5, *ns*) or weak (t(19)=1.4, *ns*) predicates.

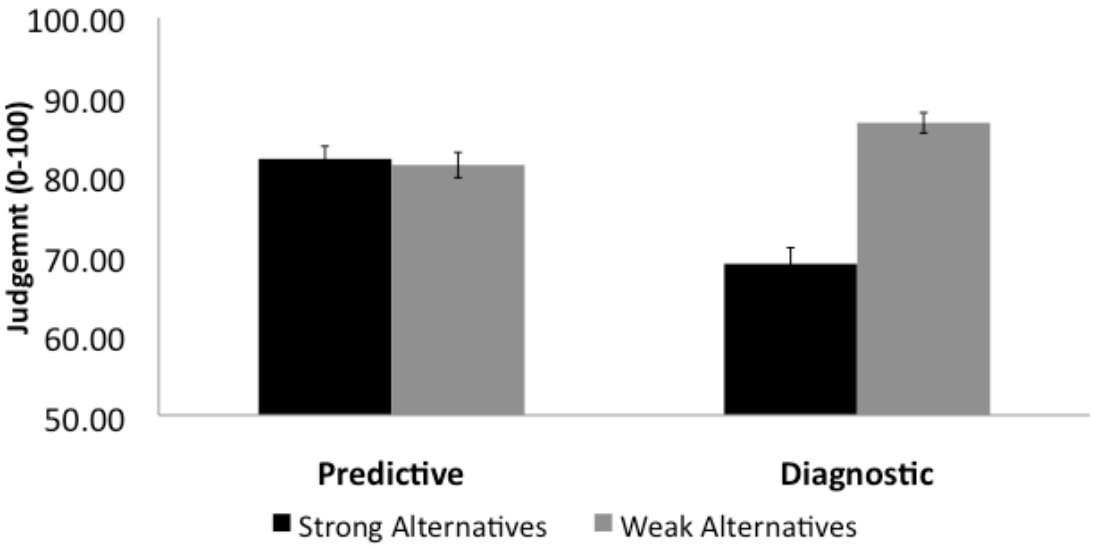

Figure 2.7: Mean Predictive and Diagnostic judgments and standard errors for the strong and weak alternatives conditions of Experiment 4.

2.6.3 Discussion

The results of Experiment 4 corroborated Experiments 1 and 2 using less vague frequency formatted questions. This supports the hypothesis that the failure to consider alternatives in predictive judgment is not driven by participants interpreting *P* questions as requesting W_c but rather by neglect of alternative causes. W_c questions were rated as slightly lower than the *P* questions. Though this difference was not significant, one might ask whether it might have become so with additional data. While this is logically possible, the small difference is not sensitive to the strong/weak manipulation suggesting that it does not represent even partial consideration of alternatives.

2.7 General Discussion

I have provided a normative analysis of inductive reasoning about transmitted predicates and reported four experiments testing how people's predictive and diagnostic inferences compare to the analysis. In Experiments 1 and 2 I collected model parameters along with predictive and diagnostic judgments, allowing me to fit the model. Participants were sensitive to alternative strength when reasoning diagnostically but neglected alternatives when reasoning predictively. Participants' diagnostic reasoning was also sensitive to the other factors highlighted by the normatve analysis, causal power and prior probability. Experiment 3 provided further evidence for neglect in the predictive direction by demonstrating that mentioning alternatives leads participants to give higher *P* judgments. The fact that participants did not raise their W_c judgments speaks against pragmatic explanations based on participants treating both questions the same. Experiment 4 replicated Experiments 1 and 2 using questions that asked about the number of cases to

which the predicate applied, providing further evidence that the effect is not due to a misinterpretation of the questions.

2.7.1 Comparison to Other Models of Property Induction

Similarity-Based Models

Similarity-based approaches such as the similarity-coverage model (Osherson et al., 1990) and the feature-based model (Sloman, 1993) propose that inductive strength is a function of the similarity between the categories in the argument and make no differential predictions based on predicate differences. These models do sometimes predict asymmetries in arguments, but these asymmetries are driven by the typicality or distinctiveness of the categories and not by the causal structures suggested by predicates. The manipulations in my experiments kept categories constant while varying the predicate and hence similarity-based models cannot account for my results.

Bayesian Models

Likewise, Bayesian models of the type proposed by Heit (1998) and Tenenbaum and Griffiths (2001; also see Sanjana & Tenenbaum, 2003) derive their predictions indirectly from similarity relations between categories and hence make no differential predictions based on predicate differences.

Other Bayesian models (Shafto et al., in press; Kemp and Tenenbaum, 2003) use different prior distributions based on the categorical relational structure that distinct classes of predicates bring to mind, and hence do make different predictions for different classes of predicates. For example, Shafto et al. propose a food web model to make predictions about transmitted predicates and a taxonomic model to make predictions about genetic properties. The model predicts an asymmetry favoring the predictive

direction but only for transmitted predicates. This asymmetry always holds in the networks they tested because the background transmission rate is held constant across all nodes in the network. A generalization of the model that allowed for different background rates for causes and effects could represent the manipulation of strong versus weak alternative predicates by varying this background rate across nodes and would be consistent with my normative formulation. In principle therefore, more specific information about the structure and parameters of an argument could be embodied in a Bayesian model of the type Shafto et al. propose that derived a prior distribution from a relational structure capturing beliefs about specific predicates.

Rehder (2009) proposes a property generalization model that represents the causal structure of predicate transmissions at the level of individual categories rather than at the level of relations between categories. This model is also related to my normative analysis in representing causal transmission in terms of noisy-or Bayes nets.

GAP Models

Smith, Shafir and Osherson (1993) proposed the GAP model to account for arguments about non-blank predicates that violate the predictions of similarity-based models. For instance (a) is rated as a stronger argument than (b) despite the fact that Poodles are less similar to German Shepherds than are Collies.

(a) Poodles can bite through wire
Therefore German Shepherds can bite through wire

(b) Collies can bite through wire
Therefore German Shepherds can bite through wire

The idea behind the model is that a more surprising or implausible premise increases the conditional probability of the conclusion because it leads to greater belief revision. The

fact that poodles can bite through wire is more surprising than the fact that Collies can and this leads to more change in belief about German Shepherds. Blok, Medin and Osherson (2007) further developed this idea with the *SimProb* model, according to which the conditional probability of a conclusion for a one-premise argument is:

$$P(Conclusion \mid Premise) = P(Conclusion)^{\left[\frac{1-SIM(premise,conclusion)}{1+SIM(premise,conclusion)}\right]^{1-P(premise)}} \quad (5)$$

where *SIM(premise, conclusion)* is the similarity between the premise and conclusion categories, which varies between zero and one and is maximal at one. The intuition behind the equation is that conditional probability is a joint function of the similarity of the premise and conclusion categories, and the plausibility of the premise, which is represented by 1-*P(Premise)*. Translating the *SimProb* equation to my materials I can define equations for the *SimProb* predictions for *P* and *D* as follows:

$$\begin{aligned} P_{simprob} &= P(Effect \mid Cause) = P(Effect)^{\left[\frac{1-SIM(Cause,Effect)}{1+SIM(Cause,Effect)}\right]^{1-P(Cause)}} \\ &= P_e^{\left[\frac{1-SIM(Cause,Effect)}{1+SIM(Cause,Effect)}\right]^{P_c}} \end{aligned} \quad (6)$$

$$\begin{aligned} D_{simprob} &= P(Cause \mid Effect) = P(Cause)^{\left[\frac{1-SIM(Effect,Cause)}{1+SIM(Effect,Cause)}\right]^{1-P(Effect)}} \\ &= P_c^{\left[\frac{1-SIM(Effect,Cause)}{1+SIM(Effect,Cause)}\right]^{1-P_e}} \end{aligned} \quad (7)$$

In Experiment 2 I collected judgments of P_c and P_e so the only thing missing for fitting the *SimProb* equations is the similarity of the cause and effect categories in the arguments. I did not collect similarity judgments in the experiments, but the design of the study introduces some constraints. I held categories constant across the strong/weak and

predictive/diagnostic manipulations so the similarity judgments for categories should not vary as a function of question type or alternative strength. With the simplifying assumption that similarity is symmetric (i.e. S*IM(effect, cause) = SIM(cause, effect)*), I can apply the *SimProb* model to my data by introducing five similarity parameters, one for each category set used in the experiment.

I explored the parameter space by varying each of the parameters from .1 to .9 in increments of .1 and calculating the *SimProb* equations at each point using the mean values of P_c and P_e from Experiment 3 collapsed over participants. This resulted in model fits at 59,049 points in the space. For each point, I calculated the correlations between mean judgments of P and $P_{simprob}$ and between D and $D_{simprob}$. The mean correlation over all points for predictive judgments was 0.24 (p=0.50) and for diagnostic judgments it was 0.0033 (p=0.93). The maximal values were 0.81 (p=0.0045) and 0.68 (p=0.031) respectively. For comparison the normative model was highly correlated with both predictive judgments ($r(8)$=0.86, p=0.0015) and diagnostic judgments ($r(8)$=0.80, p=0.0054). Despite the fact that the *SimProb* model had five free parameters versus zero for the normative model, its best fits were inferior to those achieved by the normative model and on average, it was not significantly correlated with P or D.

In addition to the correlational analyses, I assessed the qualitative fit of *SimProb* to the main finding of Experiments 1 and 2, the interaction between judgment type and alternative strength. For four out of the five predictive questions in Experiment 3, P was judged higher for the strong than the weak predicate. For all five of the diagnostic questions, the weak alternatives predicate yielded higher judgments. In line with this finding the normative model predicted that four of the five strong predicates would yield

higher predictive judgments and that all five of the weak predicates would yield higher diagnostic judgments. Conversely, *SimProb* predicted on average that 4.11 of the strong predicates would yield higher diagnostic judgments. Moreover, for two categories, *SimProb* never predicted that the diagnostic judgment should be higher for the weak predicate. *SimProb* was better at matching the predictions of predictive arguments where it predicted on average that 4.33 strong predicates would be higher. In other words, *SimProb* tended to predict that strong alternative items should yield higher predictive and diagnostic judgments while participants and the normative model generated the interaction.

I also tested *SimProb* by asking 12 people for the similarity parameters and using the mean of each parameter to fit the model. The mean values were 0.46 for *SIM(Mother, Newborn)*, 0.44 for *SIM(Coach, Team),* 0.43 for *SIM(Apple Slices, Apple Pie),* 0.40 for *SIM(Music at Party, Party),* and 0.45 for *SIM(Engine, Honda Accord)*. The analysis yielded non-significant correlations to *P* ($r(8)$=0.37, p=0.29) and to *D* ($r(8)$=0.14, p=0.69), and both correlations were significantly lower than those of the normative model (*P*: p=0.0087; *D:* p=0.0058). *SimProb* also predicted that all five of the strong alternatives items should yield higher predictive judgments and higher diagnostic judgments than the weak items, again inconsistent with the interaction.

In summary, *SimProb* failed to capture the qualitative result from Experiments 1 and 2, the interaction between question type and alternative strength. It also could not match the quantitative performance of the normative model, even with the advantage of five free parameters. It should be noted that *SimProb* is not aimed at modeling transmissions between premise and conclusion categories and as such it is not surprising

that it cannot match the data. The results also do not imply that *SimProb* fails to capture reasoning about arguments of the type to which Blok, Medin and Osherson (2007) apply it. The analyses simply show that reasoning about transmission predicates that draws on causal structure knowledge cannot be explained by premise plausibility.

2.7.2 Causal Asymmetry

I began the paper with a discussion of two studies in the inductive reasoning literature, Tversky and Kahneman's (1982) proposal that the ease of reasoning from causes to effects leads to a bias to overestimate predictive judgments relative to diagnostic judgments, and Medin et al.'s (2003) relevance framework that predicts a causal asymmetry for a similar reason. My analysis is contrary to Tversky and Kahneman's findings. They chose situations with identical predictive and diagnostic probabilities and showed that people rated the predictive direction as higher. My analysis allowed me to assess normativity for a wider range of situations because it predicts the relative strength of *P* and *D* for all parameter values. In general, people's bias was in the opposite direction to Tversky and Kahneman's proposal. Predictive judgments were systematically under-estimated while diagnostic judgments were unbiased[6]. What could explain the divergence between their study and mine? One possibility is a difference of methodology. Tversky and Kahneman asked their participants to choose which of the two probabilities were higher while I asked people to estimate probabilities for individual questions.

To assess this idea I attempted to replicate one of Tversky and Kahneman's (1982) examples with my procedure. I asked 20 people to estimate the likelihood that a daughter has blue eyes given that her mother does, and another 20 to estimate the likelihood that a mother has blue eyes given that her daughter does[7]. I found no evidence

for a bias in the predictive direction. *D* was actually rated higher than *P* (M_D=49.9; M_P=42.5) but this difference was not significant ($t(38)=1.02$, $p=0.31$). In other words, Tversky and Kahneman's finding obtained by asking people to judge which probability is higher does not generalize to a direct probability judgment task.

This work also provides a different way to understand the causal asymmetry reported by Medin et al. (2003). It suggests that psychological principles like ease of reasoning are not necessary to explain the phenomenon because predictive judgments about transmissions should usually be stronger than diagnostic ones. Based on 10,000 samples taken from the joint uniform distribution over all three parameters, I found that *P* is greater than *D* in 65% of the parameter space. This implies that predictive judgments should tend to be higher than diagnostic ones and suggests that the asymmetry reported by Medin et al. may be a result of differences in the evidential value of the premises in the two directions of reasoning. Even though my results show that on balance people underestimate predictive judgments relative to diagnostic judgments, the informational asymmetry in the materials may have been sufficient to yield an asymmetry in the predictive direction. Further support for this idea comes from Shafto et al. (in press) who show that causal asymmetry arises naturally when inductive judgments are derived from a food web via a rational Bayesian comparison.

2.7.3 Conclusions

These data suggest that there is no "causal asymmetry" in that predictive judgments are not judged higher than diagnostic ones, all else being equal. But I have identified a different kind of asymmetry, an asymmetry in the extent to which people are sensitive to the strength of alternatives in the two directions of reasoning. Diagnostic judgments are

inherently comparative in the sense that they are, in part, a measure of how likely the target cause was to have brought about the effect relative to other causes. In the most direct kind of diagnostic task (i.e., a judgment of the conditional probability of a cause given an effect) this comparison comes naturally and leads to responses that closely approximate the normative calculation. In contrast, people neglect alternatives when generating predictive probabilities and hence underestimate the likelihood of effects, even though they take alternatives into account if you remind them. In some ways this is a paradoxical result. Diagnostic reasoning is more complex in that it requires considering all three factors -- prior probability, causal power and alternatives – while predictive reasoning is a simpler function of two of them. This suggests that the stumbling block to good inductive reasoning is not the complexity of the required computations. People have the capacity to make good judgments when they consider the right factors, but they fail to take into account all that they should. This chapter has identified one factor that determines whether people will use the necessary information: the causal directionality of inference.

3. Neglect of Alternative Causes[8]

3.1 Introduction

In Chapter 2 I performed a normative analysis of predictive and diagnostic reasoning about causal transmissions (e.g., the likelihood a baby has a drug addiction given her or his mother does vs. the likelihood a mother has a drug addiction given her baby does). To test the analysis, I collected judgments of people's underlying beliefs about the causal scenarios (e.g., base rates, causal strength, and strength of alternatives) and compared people's predictive and diagnostic judgments with those implied by the analysis based on the beliefs. The analysis predicted people's diagnostic judgments but overestimated their predictive judgments. This is indirect evidence that people neglect alternatives, but only in the predictive direction.

In this chapter, I take a more direct approach to assessing the role of alternative causes in the two directions of inference. My method is to compare standard predictive and diagnostic judgments: those in which alternative causes are implicit (*full conditionals*) with those in which participants are told that no alternative causes are present (*no-alternative conditionals*). The design is depicted in Table 3.1. Excepting unusual circumstances (Pearl, 1988), alternative causes increase the likelihood of the effect. Therefore, full-conditional probabilities should be judged as higher than no-alternative conditionals. Conversely, in diagnostic reasoning, alternative causes compete to explain the effect and therefore should yield lower probability judgments (often called *discounting* or *explaining away*). Full conditionals should therefore be judged as less likely than no-alternative conditionals. If participants neglect alternatives in prediction, but not in diagnosis, then their judgments of full and no-alternative predictive

conditionals should be the same, but full diagnostic conditionals should be judged as less likely than no-alternative conditionals.

Table 3.1: The Design of Experiments 1–3

	Conditional	
Judgment	Full	No alternative
Predictive	*P(effect|cause)*	*P(effect|cause, no alternative causes)*
Diagnostic	*P(cause|effect)*	*P(cause|effect, no alternative causes)*

Experiment 1 tests the neglect hypothesis in an expert population: mental health practitioners reasoning about a case study. Experiment 2 tests inferences about people's goals and means to achieving those goals, extending existing research on goal shielding (Shah, Friedman, & Kruglanski, 2002). Experiment 3 manipulates strength of alternatives in arguments involving causal transmission. Experiments 2 and 3 allow me to assess how judgments about full and no-alternative conditionals vary with the strength of alternatives.

3.2 Experiment 1

Medical judgment suffers from the same biases as those observed in everyday judgment (Bornstein & Emler, 2001). One purported source of error is the neglect of alternative causes (e.g., diseases or other medical conditions) when clinicians are called on to make prognoses or diagnoses. Experiment 1 tested whether mental health practitioners would neglect alternatives when making a predictive (prognostic) as opposed to a diagnostic medical judgment.

3.2.1 Method

Two hundred sixty-five mental health practitioners participated as part of a psychopharmacology review course offered by the Massachusetts General Hospital Psychiatry Academy (70% MD; 51% female and 49% male). Participation was voluntary; 56% of course attendees completed the experiment. The participants were assigned alphabetically to one of two groups. The predictive group answered two predictive questions: one full conditional and one no-alternative conditional. The diagnostic group answered two diagnostic questions: one full conditional and one no-alternative conditional. The questions are shown in Table 3.2. Responses were made on a 10-point scale, ranging from 1, *least likely*, to 10, *most likely*. The full-conditional question was always asked first and was completed on the first day of the course. The no-alternative question was presented the next day.

Table 3.2: Questions From Experiment 1

	Conditional	
Judgment	Full	No alternative
Predictive	Ms. Y is a 32-year-old female who has been diagnosed with depression. Please indicate on the scale below from 1 to 10 (1 being the least likely and 10 being the most likely) the likelihood that she presents with lethargy.	Ms. Y is a 32-year-old female who has been diagnosed with depression. A complete diagnostic workup reveals that she has not been diagnosed with any other medical or psychiatric disorder that would cause lethargy. Please indicate on the scale below from 1 to 10 (1 being the least likely and 10 being the most likely) the likelihood that she presents with lethargy.
Diagnostic	Ms. Y is a 32-year-old female who presented with lethargy. Please indicate on the scale below from 1 to 10 (1 being the least likely and 10 being the most likely) the likelihood that she has been diagnosed with depression.	Ms. Y is a 32-year-old female who presented with lethargy. Please indicate on the scale below from 1 to 10 (1 being the least likely and 10 being the most likely) the likelihood that she has been diagnosed with depression given that a complete diagnostic workup revealed that she has not been diagnosed with any other medical or psychiatric disorder that would cause lethargy.

3.2.2 Results and Discussion

Mean judgments for the predictive and diagnostic questions are shown in Figure 3.1. To analyze the data, I performed a 2 (direction of inference: predictive/diagnostic) × 2 (conditional type: full/no-alternative) analysis of variance with repeated measures on the conditional type factor. The analysis revealed a significant interaction between direction of inference and conditional type, $F(1, 263) = 16.5, p < .0001, \eta^2 = .06$, as predicted. There was also a main effect of direction of inference, $F(1, 263) = 9.1, p < .01, \eta^2 = .03$, and conditional type, $F(1, 263) = 12.1, p < .001, \eta^2 = .04$. Follow-up planned comparisons revealed a significant difference between full ($M = 5.9$) and no-alternative ($M = 6.7$) diagnostic conditionals, $t(129) = 4.9, p < .0001$, Cohen's $d = 1.1$, but not

between predictive full ($M = 6.9$) and no-alternative ($M = 6.8$) conditionals, $t(134) = 0.5$, $p > .6$, Cohen's $d = 0.04$.

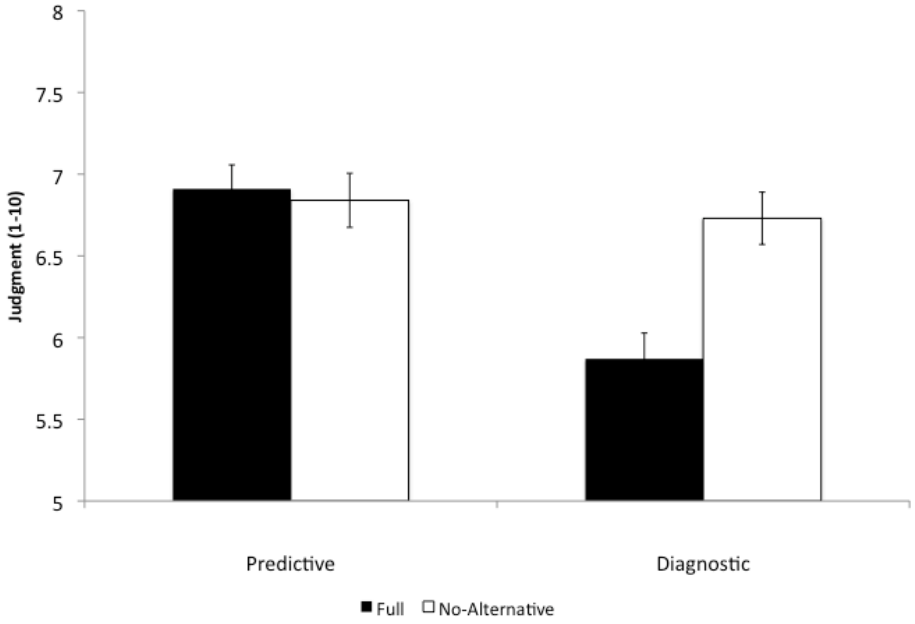

Figure 3.1: Mean likelihood ratings as a function of type of judgment (predictive or diagnostic) and type of conditional (full or no alternative) in Experiment 1. Responses were made on a 10-point scale, ranging from 1, least likely, to 10, most likely. Error bars represent standard errors.

Predictive judgments were insensitive to the absence of alternative causes. Ratings for diagnostic judgments were higher when alternatives were absent, as they should be. The results support the conclusion that the medical professionals neglected alternatives when reasoning from disease to symptom but took them into account to make a diagnosis.

3.3 Experiment 2

Experiment 1 established neglect of alternative causes in prediction but not diagnosis in an expert population reasoning about a single case. In Experiment 2, I sought to generalize the phenomenon to lay reasoning about multiple scenarios from a novel domain. One role of predictive and diagnostic reasoning is to inform choices about how

to achieve goals. Evaluating a plan of action requires predicting the likelihood of success. Evaluating actions in retrospect requires diagnosing whether they were important for having achieved the goal. Shah, Friedman, and Kruglanski (2002) showed that thinking about one means to achieving a goal reduces thinking about or pursuing alternative means in a variety of tasks. This is reminiscent of the neglect of alternatives in prediction. This implies that the effect may not be domain-specific but rather due to a more general causal reasoning process that applies across multiple domains. If this proposal is correct, then reasoning about goal schemata should evidence the same pattern and the neglect of alternative means should be mitigated when participants are to judge the diagnostic likelihood of a means given that a goal has been achieved.

A secondary objective of Experiment 2 was to assess how predictive and diagnostic judgments very with the strength of alternatives. The causal model analysis from Chapter 1 suggests that diagnostic full judgments should decrease as alternative strength increases but that diagnostic no-alternative judgments should be at ceiling regardless of alternative strength. Neglect of alternatives in prediction suggests that strength of alternatives should have no effect on full or no-alternative predictive judgments.

3.3.1 Method

Seventy-five Brown University students were recruited on campus and participated voluntarily. They were randomly divided into five groups. Groups 1 and 2 gave full-conditional judgments, Groups 3 and 4 gave no-alternative judgments, and Group 5 rated the strength of alternatives. I generated questions for eight goal schemata. Each group answered one question about each schema, and the questions were split so that no

participant saw both the predictive and diagnostic questions for a given schema. Thus, for each predictive question that Group 1 answered, Group 2 answered the corresponding diagnostic question and vice versa (and likewise for Groups 3 and 4). The presentation order of the schemata was determined randomly and was the same across all groups. The five questions for one of the schemata are shown in Table 3.3. The additional schemata can be viewed in Appendix B. The eight questions were displayed on a single page with instructions at the top, and the questionnaire took 5 to 10 min to complete.

Table 3.3 Example Questions From Experiment 2

Question type	Wording
Full predictive	Imagine you exercise hard in April. How likely is it that you weigh less in May?
No alternative predictive	Imagine you exercise hard in April. You don't have the opportunity to do anything else to lose weight besides exercising hard. How likely is it that you weigh less in May?
Full diagnostic	Imagine you weigh less in May than April. How likely is it that you exercised hard in April?
No alternative diagnostic	Imagine you weigh less in May than April. You didn't have the opportunity to do anything else to lose weight besides exercising hard. How likely is it that you exercised hard in April?
Strength of alternatives	Imagine you don't exercise hard in April. How likely is it that you weigh less in May?

3.3.2 Results and Discussion

Mean judgments for the predictive and diagnostic questions are shown in Figure 3.2a. A 2 (direction of inference: predictive/diagnostic) × 2 (condition type: full/no alternative) analysis of variance revealed a significant interaction between direction of inference and condition type, $F(1, 58) = 22.4, p < .0001, \eta^2 = .3$, as predicted by the neglect hypothesis. There were also main effects of direction of inference, $F(1, 58) = 10.6, p < .01, \eta^2 = .2$, and condition type, $F(1, 58) = 24.3, p < .0001, \eta^2 = .3$. Planned comparisons revealed a

significant difference between full (M = 54.7) and no-alternative (M = 83.3) diagnostic conditionals, $t(58) = 7.0$, $p < .0001$, Cohen's $d = 1.8$, but none for full (M = 59.2) or no-alternative (M = 58.8) predictive conditionals, $t(58) < 0.08$, $p > .9$, Cohen's $d = .02$.

I performed a median split of the schemata into strong versus weak alternatives on the basis of the strength-of-alternatives ratings. Mean conditional probability judgments for each group are shown in Figure 3.2b. To assess the effect of strength of alternatives on full and no-alternative judgments, I compared responses to strong and weak items separately for each type of question. Strong alternatives yielded lower diagnostic full-conditional judgments than weak alternatives, $t(56) = 4.3$, $p < .0001$, Cohen's $d = 1.1$. None of the other groups showed differences across the strong/weak factor.

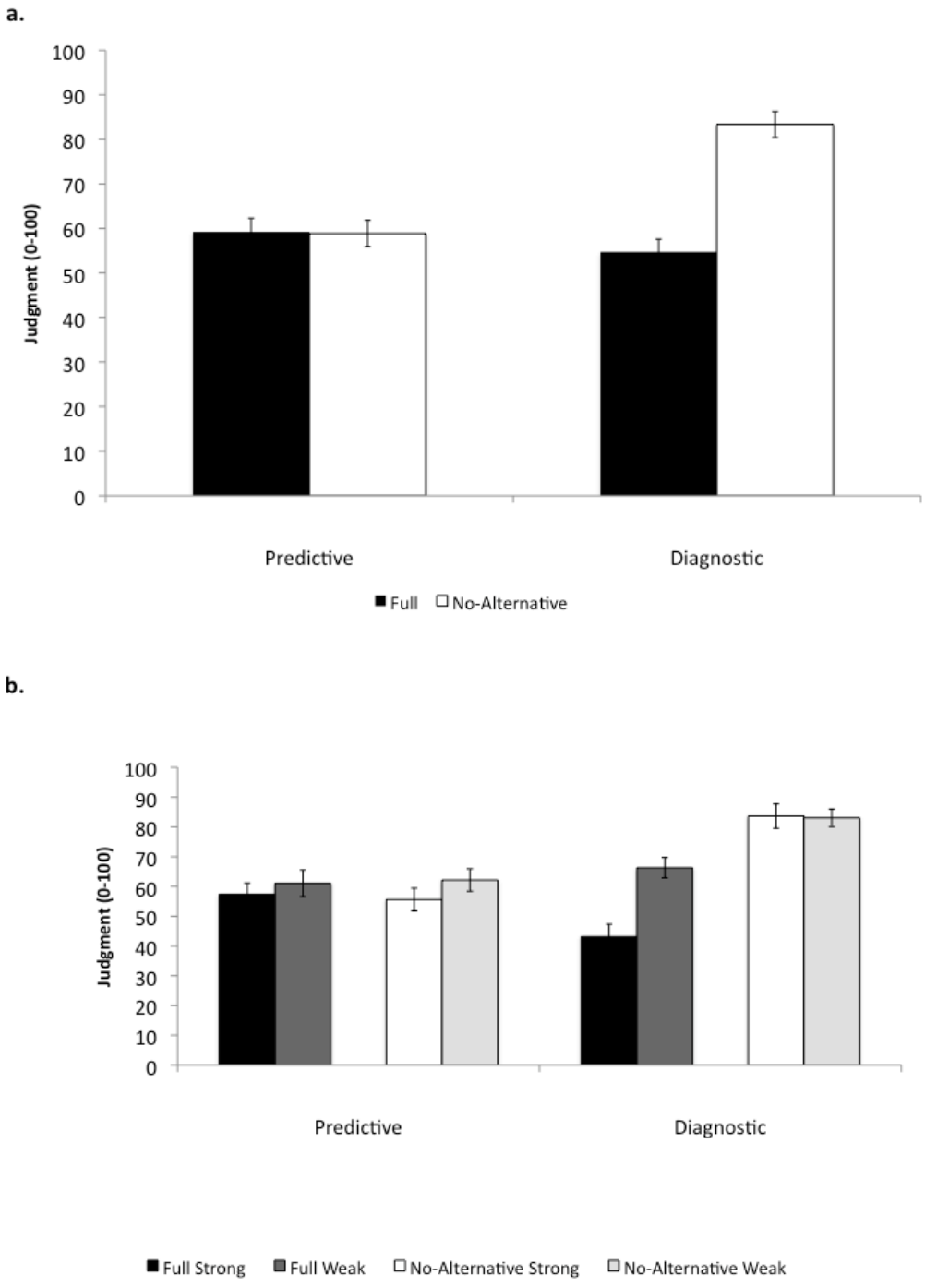

Figure 3.2: Mean likelihood ratings as a function of (a) type of judgment (predictive or diagnostic) and type of conditional (full or no-alternative) and (b) type of judgment, type of conditional, and type of alternative schemata (strong or weak) in Experiment 2. Error bars represent standard errors.

As in Experiment 1, the presence of alternatives influenced only diagnostic judgments and did so appropriately: Strong alternatives lowered full diagnostic conditional judgments to a greater degree than weak alternatives, but the strength of alternatives had no effect on no-alternative diagnostic judgments. Predictive judgments were insensitive to both the strength and even absence of alternatives. The results again suggested that people neglect alternatives in the predictive direction but treat alternatives appropriately when reasoning diagnostically.[9]

3.4 Experiment 3

Experiment 3 was designed to replicate and extend the results of Experiment 2 to causal transmission arguments of the type tested in Chapter 2. Causal transmission arguments are a special case of property projection across categories. In philosophy, property projection provides the prototypical illustration of the 'riddle of induction' (Goodman, 1955). As such, they have served as the test case for many theories of inductive reasoning in psychology (Rips, 1975). If the phenomenon is established with these arguments, it suggests that one way that people solve the property projection problem is by representing the causal relations between categories and using the same causal reasoning processes as when making other kinds of predictive and diagnostic inferences.
Another benefit of these materials is that they manipulate strength of alternatives, allowing further validation of the pattern of neglect in Experiment 2. Causal transmission arguments are a

3.4.1 Method

Sixty-three Brown University students participated for class credit or were paid $8 per hour. They were divided into four groups. Groups 1 and 2 answered the full-conditional

questions, and Groups 3 and 4 answered the no-alternative conditional questions. Each question referred to a causal transmission in which a predicate was transmitted from a cause category to an effect category. For each set of categories, I used two predicates, one that implied strong alternative causes and one that implied weak alternative causes. In Chapter 2 I verified that the strong predicates yielded higher alternative-strength judgments than did the weak predicates. As in Experiment 2, no participant saw the predictive and diagnostic questions for a particular predicate. I used 10 sets of categories and two predicates per set. Each participant therefore answered 20 questions.

The four questions for a weak and strong version of an example argument are shown in Table 3.4. The additional categories and predicates can be viewed in Appendix B. The procedure was identical to Experiment 2 except that the questionnaire was completed on a computer in a lab.

Table 3.4: Example Questions From Experiment 3

	Predicate	
Question type	Strong alternative	Weak alternative
Full predictive	The coach of a high school football team is highly motivated. How likely is it that his team is highly motivated?	The coach of a high school football team knows a complicated play. How likely is it that his team knows a complicated play?
No-alternative predictive	The coach of a high school football team is highly motivated. Imagine a situation in which there are no other possible causes of the team being motivated except for the coach. How likely is it that the team is highly motivated?	The coach of a high school football team knows a complicated play. Imagine a situation in which there are no other possible causes of the team knowing a complicated play, except for the coach teaching it to them. How likely is it that the team knows a complicated play?
Full diagnostic	A high school football team is highly motivated. How likely is it that their coach is highly motivated?	A high school football team knows a complicated play. How likely is it that their coach knows a complicated play?
No-alternative diagnostic	A high school football team is highly motivated. Imagine a situation in which there are no other possible causes of the team being motivated except for the coach. How likely is it that the coach is highly motivated?	A high school football team knows a complicated play. Imagine a situation in which there are no other possible causes of the team knowing a complicated play, except for the coach teaching it to them. How likely is it that the coach knows a complicated play?

3.4.2 Results and Discussion

Mean judgments for the predictive and diagnostic questions are shown in Figure 3.3a. A 2 (direction of inference: predictive/diagnostic) × 2 (conditional type: full/no alternative) analysis of variance revealed a significant interaction between direction of inference and conditional type, $F(1, 61) = 62.3, p < .0001, \eta^2 = .5$, and main effects of direction of inference, $F(1, 61) = 18.0, p < .0001, \eta^2 = .2$, and of conditional type, $F(1, 61) = 24.9, p <$

.0001, $\eta^2 = .3$. Planned comparisons revealed a significant difference between full (M = 69.3) and no-alternative (M = 93.9) diagnostic conditionals, $t(61) = 8.4$, $p < .0001$, Cohen's $d = 2.2$, but no difference for full (M = 74.7) or no-alternative (M = 5.8) predictive conditionals, $t(58) = 0.4$, $p > .7$, Cohen's $d = 0.09$. Mean responses for the strong and weak predicates are shown in Figure 3.3b. As in Experiment 2, alternative strength had a significant effect only on full diagnostic judgments, $t(64) = 7.8$, $p < .0001$, Cohen's $d = 2.0$.

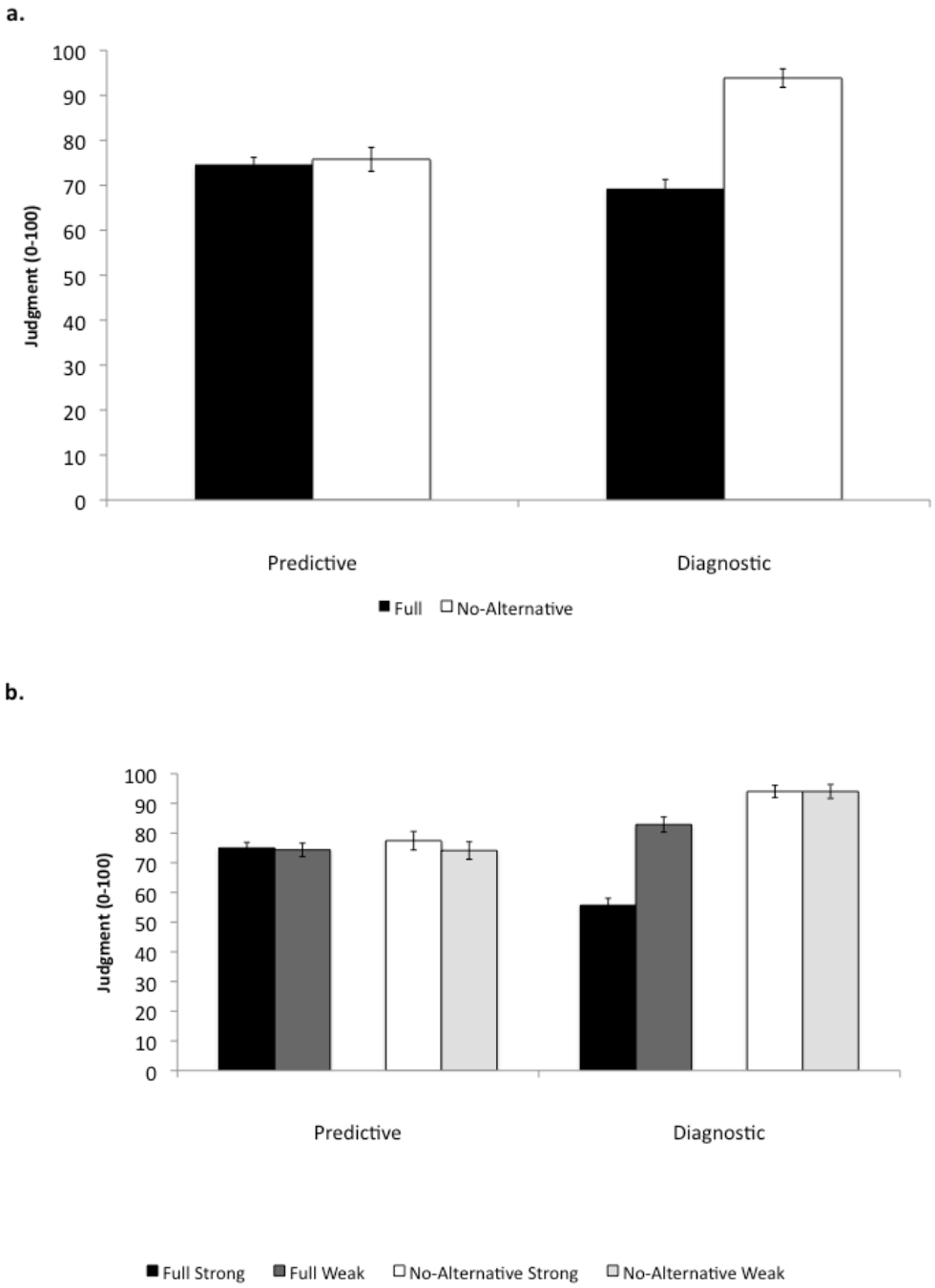

Figure 3.3: Mean likelihood ratings as a function of (a) type of judgment (predictive or diagnostic) and type of conditional (full or no-alternative) and (b) type of judgment, type of conditional, and type of predicate (strong or weak) in Experiment 3. Error bars represent standard errors.

The pattern of results in Experiment 3 was identical to Experiment 2: The explicit absence of an alternative cause affected diagnostic but not predictive judgments, and strength of alternatives affected only full diagnostic judgments.

3.5. General Discussion

Whether experts reasoning about psychopathology or undergraduates reasoning about their goals and actions or causal transmissions, people neglected alternative causes when making predictive-likelihood judgments but were sensitive to them when reasoning diagnostically.

3.5.1 Alternative explanations

One might argue that the pattern of results reflects the special status of no-alternative diagnostic judgments. These probabilities should be rated very highly, equal to 1 or close to it, making the difference between full and no-alternative diagnostic judgments obvious. Conversely, the difference between no-alternative and full predictive judgments is subtler because neither takes a value at the end of the probability scale. This interpretation predicts the high likelihood ratings in the no-alternative diagnostic condition relative to other judgments, but not the effects of alternative strength in Experiments 2 and 3. In diagnostic reasoning, participants were sensitive not just to the presence or absence of alternatives, but to the degree of alternative strength. Predictive judgments did not vary with alternative strength.

Another potential explanation for the results is that people neglected alternatives in the predictive direction because of pragmatic considerations. Do people interpret full conditionals as containing an implicature to ignore unmentioned causes in the predictive direction only? I tried to avoid such implicatures by choosing wordings that lent

themselves more naturally to the intended interpretation: the full conditional. For example, in Experiment 1, participants were told that Ms. Y was diagnosed with depression and then were asked to judge the likelihood of her presenting with lethargy. Admittedly, it remains a logical possibility that people interpreted this as a request to judge the probability that the patient presents with lethargy that is due to depression and not any other cause, but I find this interpretation unlikely, especially given the results of Experiments 3 and 4 in Chapter 2.

3.5.2 Potential mechanisms

A more complete explanation for the divergence between predictive and diagnostic reasoning emerges from a consideration of the demands imposed by the two directions of reasoning. People make predictions by simulating the mechanisms that produce predicted states from specific causes (Hagmayer & Waldmann, 2000; Kahneman & Tversky, 1982), and people tend to simulate one or a small number of mechanisms for a particular outcome (Dougherty, Gettys, & Thomas, 1997). It is reasonable to start with the cause that is picked out by the current argument or situation. Generating novel explanations is difficult because of the vast number of potentially relevant factors (Josephson & Josephson, 1984; Peirce, 1931). Diagnostic-likelihood judgments, however, demand comparing the cause at hand with alternative possible causes; engaging in explanation is unavoidable. The presence of an explanatory process may also be why people make predictive judgments with greater confidence than diagnostic ones (Tversky & Kahneman, 1982).

3.5.3 Implications

These findings are at odds with the claim that predictions are positively biased relative to diagnoses (Medin et al., 2003; Tversky & Kahneman, 1982). Instead I found that predictions are underestimated due to the neglect of alternatives. The effect reported by Tversky and Kahneman is based on only a single question, and those reported by Medin et al. may be driven by strong alternative causes lowering diagnostic judgments and not by bias (for more details, see Chapter 2).

Neglect of alternatives in predictive-likelihood judgments implies an undue optimism in the case of medical prognoses (or pessimism regarding the success of treatments) and undue pessimism in the case of planning and goal pursuit. For example, a graduate student thinking about future job prospects in the context of his or her current research neglects the effects of future research.

In the light of research showing neglect of alternatives in some diagnostic situations (Doherty et al., 1996; Fischhoff, Slovic, & Lichtenstein, 1978), the consideration of alternatives in diagnostic-likelihood judgments is at least as surprising as the neglect in predictive ones. It is notoriously difficult to get people to consider alternative hypotheses. One debiasing strategy is to get them to consider the opposite (Lord, Lepper, & Preston, 1984). My work suggests that getting people to consider alternatives may be facilitated by having them explicitly judge how likely their hypothesis is given the evidence, especially when the hypothesis can be construed as a potential cause of the evidence. People apparently are already equipped to consider alternative causes under these conditions.

4. The Weak Evidence Effect[10]

4.1 Introduction

Deciding on a course of action often requires a prediction about the future. For instance, deciding whether to support a public policy initiative depends on the state of affairs likely to obtain if the policy is adopted. Will the economy improve if a stimulus bill is passed? Will a war-torn country eventually achieve political stability if troops are committed? Across a wide variety of judgment and reasoning tasks people tend to be myopic, basing judgments primarily on whatever happens to be in the focus of attention while neglecting relevant alternatives (Dawes, 2001). Neglect of alternative causes is robust when people are making predictions though not for diagnostic judgments (Chapters 2 and 3). As the examples illustrate, predictions about outcomes are often made in the context of particular causes for those outcomes. If the focal cause dominates judgment to the exclusion of other potential causes, serious predictive error can ensue.

To understand why such undue focus might prove insidious, consider someone who has purchased a gallon of milk and put it in the refrigerator. If the power were to subsequently go out for a short time, she might consider the short power outage to be unlikely to cause spoilage. If she were to focus unduly on that weak relation when predicting whether the milk is indeed spoiled, this could lead her to judge the probability that milk is spoiled a week later to be *lower* than if the power had not gone out. If she also believes that the short power outage increases the likelihood of spoilage (albeit only by a little bit) she is in a paradoxical situation; on one hand she believes short power outages increase the likelihood of spoilage. On the other hand, the short power outage has led to a lower assessment of the likelihood of spoilage.

The magnitude of potential error due to the neglect of alternative causes increases as the likelihood of the outcome increases and the causal strength of the focal cause – the probability that it will succeed in bringing about the outcome -- decreases. An extreme case illustrated in the spoiled milk example arises when a weak cause -- one that raises the likelihood of the effect a little bit -- actually reduces a person's confidence that the effect will occur.

Establishing such cases requires three types of judgments: i. The *conditional* likelihood of the outcome given a cause (e.g. How likely is milk spoilage given the short power outage?), ii. The *marginal* likelihood of the outcome – the likelihood of the outcome when no causes are mentioned (e.g. How likely is milk spoilage?), and iii. A *probability raising* judgment to verify that the cause in question is indeed seen as raising the likelihood of the outcome (e.g. Does the power outage raise or lower the likelihood of spoilage?). It is inconsistent to judge the conditional lower than the marginal but judge the cause as probability raising. Nonetheless, I predicted that this pattern would emerge due to the neglect of alternative causes. I refer to it as the *weak evidence effect*.

4.2 Experiment 1

Understanding the process by which people decide to support or oppose public policy initiatives is vitally important. Public policy decisions have major impacts on societies and individuals. They can be complex, difficult to understand and controversial. A good policy may fail to be adopted because a lack of public support due to a failure to understand the benefits. A bad policy might gain support for opposite reasons. As suggested above, reasoning about public policy often requires making a prediction about an outcome in the context of a particular cause or set of causes for that outcome. For

these reason I decided to explore the weak evidence effect in the domain of public policy reasoning.

I created stimuli based on four public policy themes in the public consciousness at the time of study: the economy, the climate, healthcare and the war in Afghanistan. For each theme I collected judgments of the conditional probability of an effect given a weak cause, the marginal probability of the effect, and whether the cause is probability raising. I predicted that participants would display the weak evidence effect and judge conditional likelihoods lower than marginal likelihoods while judging the causes to be probability raising. Formally, the logic is identical to the spoiled milk example above, but with materials drawn from public policy themes.

4.2.1 Methods

Conditional, marginal, and probability raising questions were created for each theme. Each conditional probability question consisted of three sentences. The first stated some background information, the second the presence of a weak cause, and the third asked the likelihood of the outcome. The conditional questions for each theme are shown in Table 4.1. Marginal questions were identical except they did not contain the second sentence. The conditional and marginal questions were split into two questionnaires such that each questionnaire had two of each and so that neither questionnaire had both the marginal and the conditional for a particular theme. Two filler questions were added so each questionnaire had six questions: two conditionals, two marginals, and two filler items. Instructions at the top of each questionnaire asked participants to judge each question on a 0 (‘impossible’) to 100 (‘definite’) scale.

Table 4.1: Stimuli From Experiment 1

Theme	*Conditional*
Economy	Approximately 10% of the US population is currently using food stamps. The Congress has recently approved a 15-cent increase in the federal minimum wage (from $7.25 to $7.40). How likely is it that the percentage of people using food stamps will be less than 9% by the beginning of 2011?
Climate	Widespread use of hybrid and electric cars could reduce worldwide carbon emissions. One bill that has passed the Senate provides a $250 tax credit for purchasing a hybrid or electric car. How likely is it that at least one fifth of the US car fleet will be hybrid or electric in 2025?
Healthcare	The infant mortality rate in the United States is currently 6.3 deaths per 1000 live births. The health care reform bill that is likely to pass into legislation includes funding for an education program to teach prospective mothers about prenatal nutrition. How likely is it that the infant mortality rate in the United States will be below 5.5 deaths per 1000 live births by 2020?
Afghanistan	The democratic government of Afghanistan is embroiled in a protracted conflict with Taliban insurgents. The European Union recently pledged 7,000 troops to provide added security in population centers. How likely is it that Afghanistan will have a stable government in 5 years?

Probability raising questions were identical to the conditional questions except that the third sentence read 'does that raise or lower the likelihood that…' Participants judged the questions on a 7-point scale with the following response options from left to right: 'it lowers it a lot,' 'it lowers it somewhat,' 'it lowers it a little,' it neither raises nor lowers it,' 'it raises it a little,' 'it raises it somewhat,' and 'it raises it a lot.' The four probability raising questions were all included on a single questionnaire with two filler items, for a total of six questions.

Fifty-one members of the Brown University community were approached on campus and participated voluntarily. They were assigned at random to one of the three questionnaires and completed it in five to ten minutes.

4.2 Results and Discussion

The means and standard errors by theme are shown in Table 4.2. As predicted, conditional judgments (Mean=33.7) were lower than marginal judgments (mean=48.7) when I collapsed over themes and compared participant means ($t(35)=5.1$, $p<.0001$, Cohen's $d=1.7$). The difference was also significant when collapsing over participants ($t(3)=3.5$, $p<.05$, Cohen's $d=4.0$).

Table 4.2: Means and Standard Errors by Theme for Experiment 1

Theme	*Conditional*	*Marginal*
Economy	22.9 (4.5)	49.5 (6.7)
Climate	43.7 (6.7)	58.5 (6.0)
Healthcare	38.4 (4.9)	50.3 (6.1)
Afghanistan	30.6 (4.1)	36.8 (5.4)
All Themes	*33.7 (2.7)*	*48.7 (3.1)*

The probability raising questions were analyzed by converting the responses to numeric values from 1 to 7, with 4 corresponding to the scale midpoint, 'it neither raises nor lowers the likelihood.' As intended, the causes were seen as probability raising. The judgments were significantly higher than the scale mid-point (Mean=4.9, $t(14)=6.4$, $p<.0001$, Cohen's $d=1.6$) and the means of all themes were above the midpoint. Of the 60 judgments across all the themes only two were below the midpoint.

I also looked at the correlation between the magnitude of difference between Marginal and Conditional judgments and the probability raising judgments. The correlation was negative but not quite significant, $r=-0.45$, $p=0.14$, two-tailed. This provides some weak support for the idea that the weaker the cause the larger the magnitude of the weak evidence effect.

4.3 Experiment 2

In Experiment 2 I wanted to replicate the weak evidence effect with a larger number of items not drawn from the public policy arena. I created materials based on 12 themes inspired by everyday events. I also wanted to assess the relation between conditional and marginal judgments and the causal power of the cause. If participants completely neglect alternatives, causal power judgments should be identical to conditional judgments. If they sometimes consider alternatives, but not sufficiently, causal power judgments should be lower than conditional judgments. I therefore collected causal power judgments in addition to conditional, marginal, and probability raising judgments. The four questions for one of the themes are shown in Table 4.3. Additional themes are shown in the Appendix C.

Table 4.3: The Four Questions for One of the Themes in Experiment 2

Question Type	*Wording*
Conditional	A man buys a half-gallon of milk on Monday. The power goes out for 30 minutes on Tuesday. How likely is it the milk is spoiled a week from Wednesday?
Marginal	A man buys a half-gallon of milk on Monday. How likely is it the milk is spoiled a week from Wednesday?
Probability Raising	A man buys a half-gallon of milk on Monday. The power goes out for 30 minutes on Tuesday. Does that raise or lower the likelihood that the milk is spoiled a week from Wednesday?
Causal Power	A man buys a half-gallon of milk on Monday. The power goes out for 30 minutes on Tuesday. How likely is it that the power going out for 30 minutes on Tuesday causes the milk to be spoiled a week from Wednesday?

4.3.1 Methods

The conditional, marginal and causal power questions were divided into three questionnaires such that each questionnaire had four of each question type. Six filler

items were added for a total of 18 questions per questionnaire. The 12 probability raising questions were all included in a fourth questionnaire, also with six filler items. The dependent measures and instructions were identical to Experiment 1.

Seventy-three Brown University undergraduates were recruited from the psychology research pool and participated in the lab for class credit. They were assigned at random to one of the four questionnaires and completed it in approximately 15 minutes.

4.3.2 Results and Discussion

As in Experiment 1, conditional judgments (Mean=46.4) were lower than marginal judgments (Mean=52.2) when collapsed over themes, $t(53)=2.2$, $p<.05$, Cohen's $d=0.6$, and when collapsed over participants, $t(11)=2.4$, $p<.05$, Cohen's $d=1.4$.

Conditional judgments were significantly higher than causal power judgments (Mean=38.1) when collapsed over themes, $t(53)=2.8$, $p<.01$, Cohen's $d=0$, and when collapsed over participants, $t(11)=2.9$, $p<.05$, Cohen's $d=1.7$. This suggests that participants did not consistently confuse conditional questions for causal power questions and that they neglected rather than completely ignored alternative causes.

Again, the causes were judged probability raising. Probability raising judgments were significantly higher than the scale mid-point of 4, Mean=5.0; $t(18)=10.4$, $p=<0.0001$, Cohen's $d=2.4$, and the means of all 12 themes were above the midpoint. Of the 228 judgments only 11 were below the midpoint.

4.4 General Discussion

Two experiments identified a weak evidence effect in predictive likelihood judgment. When participants predicted an outcome conditioned on a weak cause for that outcome,

they gave lower judgments than when predicting the outcome without any mention of the cause, despite the fact that the causes were judged to be probability raising.

In Experiment 2, causal power judgments were lower than conditional judgments. This suggests that participants did not completely neglect alternatives, but thought of them sometimes. Inductive inferences rely on retrieval from semantic memory (Dougherty, Gettys & Ogden, 1999) and this retrieval is sometimes driven by the strength of the relation between the cue and associated memory structures (Quinn & Markovitz, 1998). Alternative causes may sometimes come to mind when they are highly available. The Afghanistan item in Experiment 1 may be a case like this. The difference between conditional and marginal judgments was smaller than for the other items perhaps because the experiment was conducted days after a high profile speech on Afghanistan by President Obama. Even though the conditional question only mentioned one cause, participants may have thought of others because they were so available.

A potential pragmatic account of the weak evidence effect is that people interpret the cause mentioned in the conditional question as implying that unmentioned alternative causes should be ignored when making the prediction. I was careful to pick effects that represent verifiable, objective states of the world (e.g. the likelihood that one fifth of the U.S. fleet will be hybrid or electric in 2020) in order to make the conditional probability interpretation the simplest one. Furthermore, the difference between causal power and conditional judgments in Experiment 2 suggests that alternative causes were not always considered irrelevant. More evidence against this pragmatic account comes from the finding that people neglect alternatives even when questions are phrased in less ambiguous frequency language (Chapter 2).

4.4.1 Related Phenomena

Negative Evidence in Reasoning About a Dispute

McKenzie, Lee, and Chen (2002) have shown that when reasoning in the context of an argument with opposing sides, weak evidence of innocence will sometimes increase belief in guilt, presumably because the case offered by the defense is expected to be as strong as possible. A weak case implies an inability to amass strong evidence. Such a conclusion need not be irrational. Evaluating evidence relative to the strength of an expectation is not only reasonable but often called for. The weak evidence effect cannot be rationalized in this way because the judgments of conditional likelihood are not presented in the context of an argument that supports expectations about the strength of evidence. The task simply asked people to rate the strength of their belief given certain information. The context does not suggest that no other information could have been obtained.

Conjunction Fallacy

The conjunction fallacy occurs when a conjunction of events is judged more probable than one of the events alone. One instantiation of the phenomenon occurs when the conjunction of an outcome and its cause is judged more likely than the outcome. Kahneman and Tversky (1983) give the following example where a) is judged more likely than b):

> a) An earthquake in California sometime in 1983, causing a flood in which more than 1000 people drown.
>
> b) A massive flood somewhere in North America in 1983, in which more than 1000 people drown.

Conjunction errors like this occur when the marginal outcome probability is low and the causal power is fairly high, precisely the converse of the conditions that facilitate the weak evidence effect. Though I do not claim to provide an analysis of reasoning about conjunctions, one implication of this correspondence is that a primitive relation underlying reasoning about both joint and conditional likelihood is knowledge about causal mechanisms. Indeed, one explanation for the conjunction fallacy is that people think about the underlying mechanism connecting the cause and the effect when assessing the conjunction (Ahn & Bailenson, 1996). The weak evidence effect and the conjunction fallacy may both be due to judgments being primarily driven by the mentioned cause and its causal power to bring about the effect. When the causal power is high the judgment is also high and when the causal power is low the judgment is also low. Other considerations, such as the prior probability of the cause (in the case of the conjunction fallacy) and alternative causes (in the current case) are neglected.

Unpacking Effects

In unpacking effects (Tversky & Koehler, 1994), the judged probability of an event type (e.g., death from disease) is less than the judged probability of the event unpacked into constituents (e.g., death from heart disease or some other disease). When thinking about an event, people tend to focus on typical cases, and unpacking with atypical cases (e.g., death from pneumonia or some other disease) can reduce judgments suggesting that people neglect constituents that are not mentioned (Sloman et al., 2004). This is analogous to the present case in which mentioning an atypical (weak) cause reduces the likelihood of an effect suggesting neglect of alternative causes. The fact that both descriptions of events to be judged and evidence about a particular event are biased in

favor of what is mentioned suggests a fairly general cognitive process at work, presumably involving a tendency for judgment to focus narrowly on the contents of working memory (cf. Thomas et al., 2008).

4.4.2 Implications

Awareness of the weak evidence effect may help people avoid being persuaded when it is used as a rhetorical tool. For instance, opponents of a public policy initiative might attempt to diminish support for the initiative by focusing attention on particular aspects of it. A 15-cent increase in the minimum wage may be a beneficial part of a larger economic stimulus bill, but focusing attention on that part of the plan makes it seem unlikely to work. This is especially important for incremental changes where it is not one specific policy change that will bring about the desired effect but the combination of a large number of policy changes.

4.4.3 Conclusion

The fact that positive evidence can reduce belief is inconsistent with all well-known theories of belief updating (e.g., Anderson, 1981; Chater & Oaksford, 2008; Hogarth & Einhorn, 1992; Kemp & Tenenbaum, 2009; Thagard, 1989). It is highly consistent with what is known about human psychology however. People tend to focus on what they perceive in their immediate discourse environment and neglect other things when reasoning (Evans, Over & Handley, 2003), testing hypotheses (Doherty et al., 2003), understanding language (Keysar, Linn, & Barr, 2003), troubleshooting (Fischoff, Slovic & Lichtenstein, 1978), and inducing properties (Ross & Murphy, 1996). This kind of cognitive limitation cannot be captured by a parameter in an otherwise rational model. It

requires a model that takes into account normatively irrelevant but psychologically important information.

5. Reaction Times and Causal Conditional Reasoning[11]

5.1 Introduction

The primary purpose of this chapter is to describe reaction time data for predictive and diagnostic likelihood judgments. I explored how reaction times varied with direction of inference, number of alternative causes and number of disabling conditions. The materials used in this chapter are based on seminal work by Cummins' (1995) exploring how people reason about deductive arguments with causal content. In addition to collecting reaction times for predictive and diagnostic judgments based on her materials, I also collected judgments of causal power, prior probability and strength of alternatives and compared model predictions to Cummins' results and to the likelihood judgments (as in Chapter 2). This allowed me to corroborate Chapter 2's findings and to test the scope of the causal model conjecture -- that is, whether it also applies to causal conditional reasoning.

5.1.1 Causal Conditional Reasoning

When reasoning about deductive arguments people are biased to accept conclusions that are consistent with their beliefs and reject those that are inconsistent, regardless of argument validity (Evans, 2007). In a set of seminal papers, Cummins (1995; Cummins et al., 1991) showed that these belief biases follow systematic principles when people reason about conditional arguments with causal content. People judged the validity of four argument schemata: Modus Ponens (MP), Modus Tollens (MT), Denying the Antecedent (DA) and Affirming the Consequent (AC), though I focus on just MP and AC in this paper.

Despite MP being deductively valid and AC invalid regardless of content, Cummins predicted that for arguments where the antecedent is a cause of the consequent, acceptance rates for MP would be affected by the number of disabling conditions while AC would be affected by the number of alternative causes for the effect.

In the case of MP, thinking of a disabling condition provides a counterexample to the argument and hence may lead people to reject it. An example is given below. Cummins' predicted that (a) would be judged more acceptable than (b) because the conditional in (a) has fewer disablers; reasons why one could put fertilizer on plants and not have them grow quickly are more available than reasons why one could jump into a pool and not get wet.

(a) If Mary jumped into the swimming pool then she got wet.

Mary jumped into the swimming pool.

Therefore she got wet.

(b) If fertilizer was put on the plants then they grew quickly.

Fertilizer was put on the plants.

Therefore they grew quickly.

In the case of AC, alternative causes provide an alternative explanation for the effect and hence make the antecedent seem less necessary. For example Cummins predicted that (c) would be judged more acceptable than (d). It is hard to think of alternative causes for a gun firing besides the trigger being pulled but it is relatively easy to think of causes of wetness besides jumping into a swimming pool.

(c) If the trigger was pulled then the gun fired.

The gun fired.

Therefore the trigger was pulled.

(d) If Mary jumped into the swimming pool then she got wet.

Mary got wet

Therefore she had jumped into the swimming pool.

To test these ideas Cummins' asked one group of participants to spontaneously generate alternative causes and disabling conditions for a host of conditionals and then divided the conditionals into four groups of four conditionals each based on the number of alternatives and disablers (many alternatives, many disablers; many alternatives, few disablers; few alternatives, many disablers; few alternatives, few disablers). A different group was given the arguments based on the 16 conditionals and asked to judge the extent to which the conclusion could be drawn from the premise. Responses were on a 6 point scale from "very sure that the conclusion cannot be drawn" (-3) to "very sure that the conclusion can be drawn" (3). The results provided good support for both predictions.

5.1.2 Conditional Probability Interpretation

Following Oaksford, Chater and Larkin (2000), if the conditional schemata are interpreted in terms of conditional probability, the acceptability of MP maps onto *P(Effect|Cause)* and AC to *P(Cause|Effect)*. More concretely, judgment of the strength of the MP argument in (b) would equate to "Fertilizer was put on the plants. What is the probability they grew quickly?" Analogously, the judgment of the strength of the AC argument in (c) equates to "The gun fired. What is the probability the trigger was pulled?"

By assuming the conditional scenarios approximate a noisy-or common effect model the expressions in (1) and (2) can be derived for MP and AC respectively (as in Chapter 2).

$$MP \approx P(Effect \mid Cause) = W_c + W_a - W_c W_a \quad (1)$$

$$AC \approx P(Cause \mid Effect) = 1 - (1 - P_c)\frac{W_a}{P_c W_c + W_a - P_c W_c W_a} \quad (2)$$

As before, W_c is the causal power of the cause, the probability that the cause successfully brings about the effect (e.g. the probability that pulling the trigger causes the gun to fire), W_a is the combined strength of all alternative causes, equivalent to the probability of the effect in the absence of the cause (e.g. the probability of the gun firing given the trigger wasn't pulled) and P_c is the prior probability of the cause (e.g. the probability of the trigger being pulled).

According to the full probabilistic model MP increases with both the causal power of the cause and the strength of alternatives (because alternative causes raise the probability of the effect). However, in Chapters 2, 3 and 4 I found that people are not sensitive to the strength of alternative causes when judging predictive likelihood despite its relevance. Thus, like Cummins I predicted no effect of W_a and the model for MP is given in (3).

$$MP \approx P(Effect \mid Cause, \sim Alternatives) = W_c \quad (3)$$

5.1.3 Relation Between Cummins' Analysis and Conditional Probability Model

According to the probability model the determinants of causal inferences, and hence MP and AC acceptability, are causal power, strength of alternatives and prior probability of the cause. The number of disablers and number of alternatives are factors in the first two

parameters, respectively. Causal power is inversely related to the number of disablers. All else being equal, as the number of disablers increases, the probability that the cause fails to bring about the effect increases, corresponding to a decrease in causal power. Thus the model is consistent with the decrease in MP as number of disablers increases, as predicted and found by Cummins. However, not all disablers are equally likely or equally effective in preventing the effect. A single strong disabler could lead to a lower causal power than several weaker disablers, making number of disablers an imperfect predictor of causal power.

Similarly, the number of alternatives is a factor in strength of alternatives. All else being equal, as the number of alternatives increases so does the probability that they will bring about the effect. Therefore, the model predicts that AC will decrease with number of alternatives. As with disablers though, number of alternatives is only a partial predictor of strength of alternatives.

Despite these similarities, the model suggests that Cummins' analysis is incomplete because it only takes a single parameter into account for each judgment. The implication for MP is that its acceptability should increase with the strength of alternative causes but as discussed above I predicted no effect of alternative causes on MP. My prediction for MP only differs from Cummins in that I expected W_c to provide a better fit than number of disablers.

The model identifies three factors relevant to the acceptability of AC arguments. First, according to the model the prior probability of the cause plays an important role in diagnostic strength. For instance, a cause that is very improbable is unlikely to have occurred relative to other more likely causes and is therefore not as good an explanation

for the effect. The second factor is the overall strength of alternatives. This differs from the number of alternatives because not all alternative causes are created equal. In the probability model the strength of alternatives reflects the probability of the effect in the absence of the cause and thus is a joint function of the prior probabilities and causal powers of alternatives. For instance, even a large number of highly improbable or weak alternatives should have less effect on the judgment then a single probable, strong cause. Finally, causal power -- and hence disablers -- should have some influence on AC. All else being equal, if the causal power of the cause is higher, the cause is more likely responsible for the effect. Table 5.1 summarizes how the model predictions differ from Cummins' theory.

Table 5.1: Best Predictors for MP and AC judgments and Predictive and Diagnostic Likelihood Judgments According to Cummins (1995) and According to the model

	MP	*AC*
Cummins' Theory	No. of Disablers	No. of Alternatives
Probability Model	Causal Power (W_c)	Full Diagnostic Model
	Predictive Likelihood	*Diagnostic Likelihood*
Cummins' Theory	No Prediction	No Prediction
Probability Model	Causal Power (W_c)	Full Diagnostic Model

5.1.4 Qualitative Support for Probability Model

Some trends appear in Cummins' (1995) data that are not predicted by her theory. One is that acceptability ratings of AC for conditionals with many alternative and few disablers were lower than those with many alternatives and many disablers. Both groups had many alternatives and thus should have yielded similar AC judgments according to Cummins. The difference was replicated by De Neys, Schaeken and D'ydewalle (2002) who found

lower AC ratings for all few disabler items compared to many disabler items (with two exceptions, they used the same conditionals).

De Neys et al. (2002) proposed that when there are many disablers, they interfere with searching memory for alternatives, leading to the observed difference. A perusal of the individual conditionals suggests an alternative explanation based on the probability model. The two groups appear to vary not just in number of disablers but also in some of the factors that the probabilistic analysis says should affect diagnostic judgments. Specifically, the items that obtain low acceptability scores share the property that the cause is weak or improbable relative to the strength of alternatives (see Table 5.2). For instance, jumping into a swimming pool is improbable relative to other causes of wetness. Likewise, pouring water onto a fire is not the most common cause of a campfire going out. On the contrary, the high ratings obtain for arguments in which the cause is strong and probable relative to alternatives. There may be many alternatives for a car slowing, but braking is likely the dominant cause. Likewise, studying hard is probably the strongest cause of doing well on a test. Thus, number of alternatives may be equated across groups, but diagnostic strength is not.

Table 5.2: Mean Acceptability of AC arguments for Two Groups of Conditionals from Cummins' (1995) Exp.1

Conditional	Acceptability (-3 to 3)
Many Alternatives, Many Disablers	
If fertilizer was put on the plants, then they grew quickly	1.00
If the brake was depressed, then the car slowed down	1.00
If John studied hard, then he did well on the test	1.50
If Jenny turned on the air conditioner, then she felt cool	1.08
Many Alternatives, Few Disablers	
If Alvin read without his glasses, then he got a headache	0.75
If Mary jumped into the swimming pool, then she got wet	0.25
If the apples were ripe, then they fell from the tree	1.00
If water was poured on the campfire, then the fire went out	-0.08

Another trend unexplained by her analysis is that few alternative conditionals obtained slightly higher MP judgments than many alternative conditionals despite being equated across number of disablers. Again, the probabilistic analysis suggests why this may be so. Several of the many alternative items have somewhat low causal powers (e.g. 'if the apples were ripe then they fell from the tree') while virtually all of the few alternative items have very high causal powers (e.g. 'if the gong was struck then it sounded.'). Thus, while number of disablers was equated across groups, causal power may have varied leading to differing MP judgments

5.2 Experiment

In the Experiment I created predictive and diagnostic conditional likelihood questions based on Cummins' (1995) conditionals and collected judgments and reaction times for these questions. De Neys et al. (2002) showed that reaction times for causal conditionals basically supported Cummins' analysis. Collecting reaction times with materials phrased in conditional likelihood language allowed us to verify and extend these findings.

To test whether the probability model accounts for the causal conditional acceptability ratings I also collected judgments of the relevant parameters: the prior probability of the cause (P_c), the causal power of the cause (W_c) and the strength of alternatives (W_a) for Cummins' (1995) conditionals as in Experiments 1 and 2 in Chapter 2. Using these judgments I derived predictions with zero free parameters to which I compared Cummins' acceptability ratings. I also compared these predictions to the conditional likelihood judgments

5.2.1 Method

Participants

133 Brown University students were approached on campus and participated voluntarily or participated through the psychology research pool in return for class credit.

Design, materials and procedure

All experimental conditions used questions based on the 16 conditionals from Cummins' (1995) Experiment 1. I therefore adopted Cummins' 2 (number of alternatives; few/many) X 2 (number of disablers; few/many) design with four conditionals in each condition. Judgments were on a 0 ('impossible') to 100 ('definite') scale.

95 participants provided predictive and diagnostic likelihood judgments, fully within-participant. Each participant therefore answered 32 questions, one predictive and one diagnostic for each conditional. In order to avoid any reaction time differences due to reading time, the wordings of the questions were modified such that each had between 13 and 15 words and between 65 and 75 characters and such that the mean number of words and characters was equated across the four groups of conditionals. Examples of predictive and diagnostic questions are given in (h) and (i):

(h) John studied hard. How likely is it that he did well on the test?

(i) John did well on the test. How likely is it that he studied hard?

This part of the experiment was administered on a computer in the lab. Participants were instructed that they this was a reaction time study and were asked to go as quickly as possible while remaining accurate. For each question, participants input their answer using the number keys and hit 'return' to move to the next question. Reaction times were measured from the moment the question appeared on the screen to when the participant hit 'return'. Order of questions was randomly determined for each participant.

17 Participants provided judgments of the prior probabilities (P_c) and strength of alternatives (W_a) for the 16 conditionals. The questions were split onto two pages with all of the P_c questions on the first page and all of the W_a questions on the second page. The order of questions was randomized on each page. For each question the conditional was first stated and then the relevant likelihood question was asked. Examples of P_c and W_a questions are given in (e) and (f) respectively.

(e) If John studied hard then he did well on the test.

How likely is it that John studied hard?

(f) If John studied hard then he did well on the test.

John did not study hard. How likely is it he did well on the test?

A minority of participants interpreted the conditional statement in the P_c questions as indicating that the cause was present and therefore gave ratings of 100 for all of the P_c questions. I removed these responses from the dataset for all subsequent analyses.

An additional 21 participants judged causal power (W_c). Methods were identical except that there was just one page of questions. An example of a W_c question is given in (g).

(g) How likely is it that John studying hard for the test causes him to do well?

5.2.2 Reaction Time Results and Likelihood Judgments

Reaction Times

All statistical tests on reaction times used a log transform to normalize the data. Outliers were removed by eliminating all trials that fell more than four standard deviations above or below the participant's mean reaction time. Additionally any trial faster than 1 second was removed.

The reaction time results are depicted in Figure 5.1. The cleaned data were subjected to a 2 (direction of inference) X 2 (number of alternatives) X 2 (number of disablers) repeated measures ANOVA. There was a main effect of direction of inference; prediction (M=5.88 s) was faster than diagnosis (M=6.21 s), $F(1,95)=25.1$, $p<0.0001$. There was also a significant interaction between number of alternatives and direction of inference, $F(1,95)=4.0$, $p<0.05$. No other main effects or interactions were significant.

The interaction between strength of alternatives and direction of inference was driven by diagnostic judgments being faster for items with few alternatives (M = 6.32 s) than for items with many alternatives (M=6.09 s), $t(94)=1.95$, $p=0.05$, Cohen's $d=0.4$ (Figure 5.1a). Predictive judgments showed no difference in reaction time across the number of alternatives manipulation, $t(94)=0.61$, *ns*.

Number of disablers had no effect on reaction times for predictive judgments ($t(94)=1.16$, *ns;* Figure 5.1b*)*. To test whether differences in causal power rather than in

number of disablers might yield reaction time differences, I performed a median split of the conditionals based on the W_c judgments and compared reaction times. Predictive judgments were faster for items with high W_c (M=5.71 s) than for items with low W_c (M=6.05 s), t(94)=4.19, p<0.0001, Cohen's d=0.9 (see Figure 5.1c). Neither number of disablers nor W_c had a significant effect on diagnostic judgments.

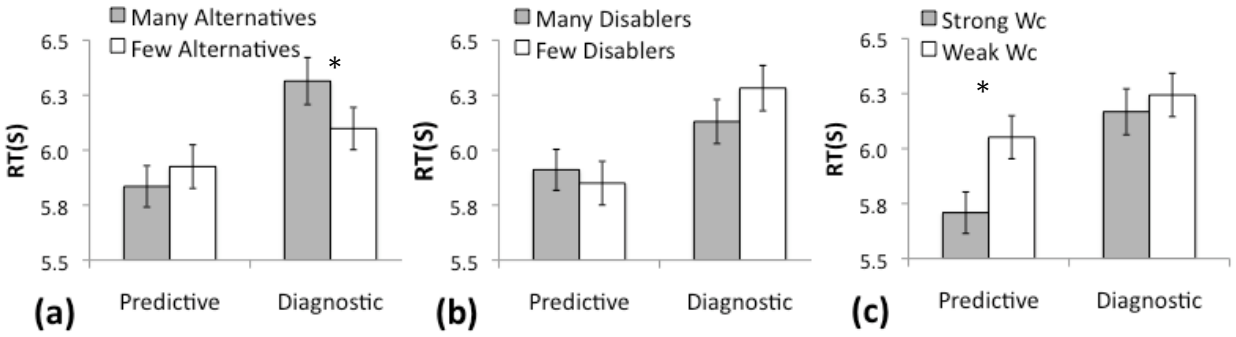

*Figure 5.1: Reaction Times for Predictive and Diagnostic Judgments by (a) number of alternatives, (b) number of disablers and (c) strength of W_c. * denotes a significant difference.*

Likelihood Judgments

Mean Predictive and Diagnostic judgments are depicted in Figure 5.2. The predictive and diagnostic likelihood judgments were subjected to a 2 (direction of inference) X 2 (number of alternatives) X 2 (number of disablers) repeated measures ANOVA. All of the main effects and two-way interactions were significant (p<0.01).

Further post hoc tests were performed on predictive and diagnostic judgments separately. Diagnostic judgments were sensitive to number of alternatives with higher judgments for the items with few alternatives (M=90.7) than for the items with many alternatives (M=57.3), t(94)=27.9, p<0.00001, Cohen's d=5.8. Diagnostic judgments also varied across number of disablers, with higher judgments for many disablers (M=78.1) than few disablers (M=70.1), t(94)=8.9, p<0.00001, Cohen's d=1.8.

As suggested by the differing W_c judgments, predictive judgments also varied across the number of alternatives; Few alternative items (M=87.8) yielded higher diagnostic judgments than those with many alternatives (M=76.3), t(94)=6.0, p<0.00001, Cohen's d=1.2. There was also a significant negative correlation between number of alternatives and predictive judgments (r=-0.49, p=.05) when correlating item means. However, there was no correlation when the effect of W_c was partialled out (r=-0.29 *ns*), suggesting that differences in W_c were responsible for weak alternative yielding higher predictive judgments.

Predictive judgments did not vary with the number of disablers (t<1, *ns*). However, I also tested whether predictive judgments varied with the strength of W_c by dividing the 16 conditionals into two equal groups based on W_c and comparing predictive judgments. As expected, conditionals with high W_c obtained higher predictive judgments (M=89.1) than those with low W_c (M=75.2), t(94)=7.0, p<0.00001, Cohen's d=1.4.

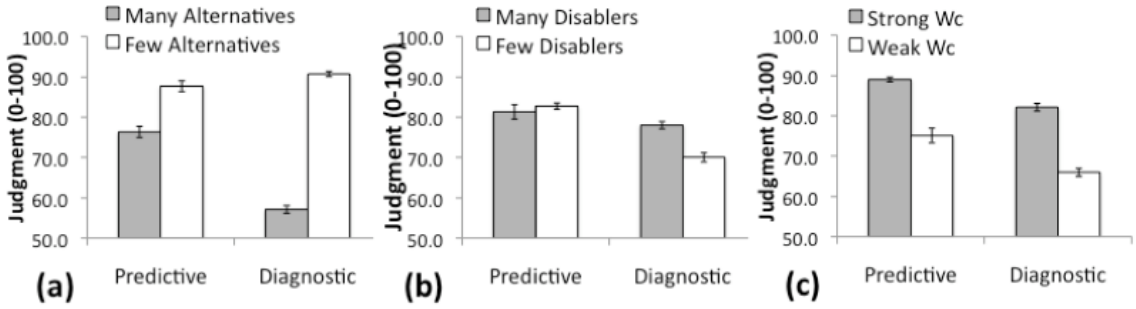

*Figure 5.2: Predictive and Diagnostic Judgments by (a) number of alternatives, (b) number of disablers and (c) strength of W_c. * denotes a significant difference.*

5.2.3 Modeling Results

Parameter Judgment Results (P, W_c and W_a)

For the following tests I collapsed over conditionals and compared participant means, using Bonferroni correction to control family-wise error rate. As expected, W_a was judged higher for many alternative items compared to few alternative items, t(16)=13.4, p<0.00001, Cohen's d=6.7, and didn't vary across few and many disablers, t(16)=1.4, *ns*.

W_c also varied across the number of alternatives manipulation; W_c was judged higher for few alternative items (M=83.4) compared to many alternative items (M=73.9), t(20)=4.8, p<0.001, Cohen's d=2.1. This was not intended by Cummins, but confirmed the intuitions about the unexplained trend in MP; weak alternative items seemed to have lower causal powers despite being equated across number of disablers. Surprisingly, W_c did not vary across the many/few disablers manipulation (t(20)=1.2, *ns*) suggesting that number of disablers and causal power were not as closely linked as I expected. The low correlation between number of disablers and W_c (r=-0.11, *ns)* also supported this conclusion. P_c did not vary across either manipulation.

Applying the Model

Simply computing Equations 2 and 3 using item means would have been inappropriate because the parameter judgments were collected between participants. I therefore used a sampling procedure to generate model predictions (as in Chapter 2, Experiment 1). For each conditional I took 10,000 samples each of W_a, P_c and W_c uniformly and randomly from participant responses, and calculated Equations 2 and 3 for each set of samples. I therefore generated 10,000 samples of each probability for each conditional and then took

the mean over samples for each conditional as the output of the model. Reruns of the model yielded only negligible differences.

Fits to AC and Diagnostic Judgments

Figure 5.3a depicts Cummins' acceptability ratings for AC on the X-axis plotted against model fits (Equation 2) on the Y-Axis for each of the 16 conditionals, along with the least squares regression line. Figure 5.3b shows diagnostic judgments plotted against model fits. The model predictions were highly correlated with both Cummins' acceptability ratings ($r=0.87$, $p<0.00001$) and the diagnostic judgments ($r=0.93$, $p<0.00001$). To test how well the model fits predicted Cummins' AC acceptability ratings, relative to the number of alternatives alone, I performed a stepwise multiple regression analysis of AC ratings using the model predictions and the number of alternatives as predictors. The model accounted for a significant amount of variance beyond what number of alternatives accounted for ($p<0.01$), and number of alternatives did not account for any unique variance ($p>0.6$). The model also accounted for variance in the diagnostic likelihood judgments that was unaccounted for by number of alternatives ($p<0.0001$).

Fits to MP and Predictive Judgments

Figure 5.3c depicts Cummins' acceptability ratings for MP plotted against model fits (equal to W_c according to Equation 3). Figure 5.3d shows predictive judgments plotted against model fits. The model was not highly correlated with MP ratings ($r=0.39$, *ns*) and number of disablers accounted for a small but significant amount of the variance beyond what the model accounted for ($p=0.041$). Conversely, the model was highly correlated with predictive judgments ($r=0.81$, $p<0.001$) and accounted for a large amount of

variance beyond number of disablers ($p<0.001$). Surprisingly, MP ratings and predictive judgments were not highly correlated, $r=0.30$, *ns*.

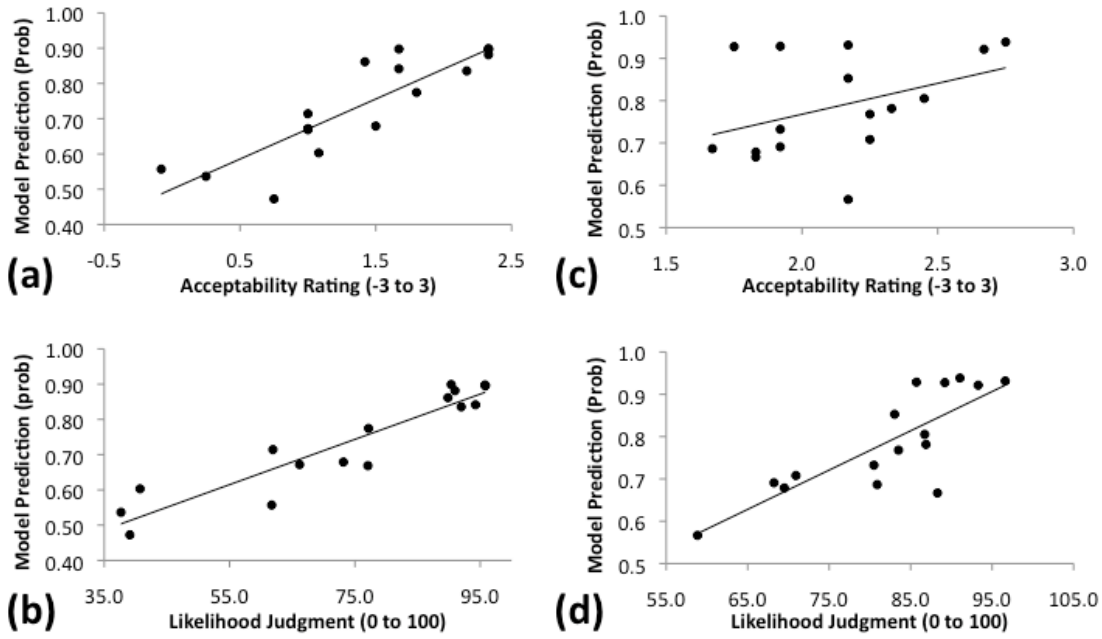

Figure 5.3: (a) Model fits against Cummins' AC acceptability ratings. (b) Model fits against diagnostic likelihood judgments. (c) Model fits against Cummins' MP acceptability ratings. (d) Model fits against predictive likelihood judgments.

5.3 General Discussion

5.3.1 Summary and Interpretation of Results

Reaction Times

The reaction time data yielded three noteworthy findings: First, predictive judgments were faster than diagnostic ones. This corroborates De Neys et al. (2002) who found that MP was faster than AC and it supports the claim that reasoning from cause to effect is easier in general than reasoning from effect to cause (Tversky & Kahneman, 1982). This difference likely reflects the time it takes to consider alternative causes and prior probability in diagnostic judgment.

Second, diagnostic judgments were faster with few alternatives. This also corroborates De Neys et al. (2002). It implies that searching for alternative causes takes time. It could also reflect the fact that when alternative causes are very weak the judgment is very high and may not require as much thought to calculate. Predictive judgments showed no reaction time differences across number of alternatives. This is more evidence that people don't think of alternatives when making predictions (Chapters 2 and 3).

Finally, I found no reaction time differences for many versus few disablers. This failed to corroborate De Neys et al. (2002) who found that MP was faster for few versus many disablers. I did however find an effect of W_c on reaction times. Prediction was faster for high versus low W_c.

Model Fits

The diagnostic model achieved very good fits to both Cummins' AC data and the diagnostic likelihood judgments with zero free parameters. It also explained more variance than the single parameter number of alternatives. This confirmed the qualitative analysis indicating that AC judgments were sensitive not just to number of alternatives, but also to the other factors in the probability model in approximately the right way. The model also accounted for the previously unexplained trend in Cummins' AC data for higher AC ratings with more disablers. Altogether, it seems that when judging AC for causal conditionals, people are actually judging the likelihood of the cause (premise) given the effect (conclusion). This implies that (at least sometimes) people use causal models to reason about deductive arguments with causal content.

The model also matched the predictive judgments closely and differences in W_c explained the previously unexplained trend in Cummins' MP judgments for higher MP judgments with fewer alternatives, a pattern that also showed up in the predictive likelihood judgments. But the model didn't match the MP data that well and in fact was slightly worse than the number of disablers at accounting for the variance. Additionally, number of disablers was a remarkably poor predictor of W_c judgments. This was surprising because I expected causal power to vary inversely with number of disablers.

5.3.2 Explaining MP

Both the model fitting and reaction times imply dissociation between how people judged MP and how they judged predictive likelihood. Predictive likelihood judgments and reaction times were explained by differences in W_c but were uncorrelated with number of disablers. Conversely, number of disablers was slightly better at accounting for Cummins' (1995) MP acceptability ratings than W_c and also yielded reaction time differences for MP in De Neys et al.'s (2002) study. This leaves three open questions: First, why is number of disablers such a poor predictor of W_c? Second, why is W_c better at accounting for predictive likelihood judgments and reaction times? Third, why is it worse at accounting for MP?

A speculative answer to the first two questions comes from the possibility that when making predictive likelihood judgments people represent causal systems in terms of their normal, common or prototypical components. If asked to list disablers they may be able to come up with a relatively large number, some of them being very uncommon or atypical. But when asked to judge causal power or make a prediction they think only of the most important disablers. The 'depressed brake' provides a good example. It's not too

hard to come up with disablers for why brakes would fail to slow a car (e.g. cut brake lines) but none of them is common. Thus, while number of disablers is relatively high, many of those disablers make a small impact on actual causal power and may have no effect on people's estimates of causal power. On this account, low causal power might still correlate with slower reaction time on the assumption that examples with a greater number of typical or high probability disablers yield lower W_c judgments, lower predictive judgments, and take longer to reason about.

This leaves the question of why W_c fails to account for MP judgments and reaction times, while number of disablers is somewhat better. I don't have a conclusive answer to this question, but suspect it may be due to people using a mixture of strategies when judging MP. In a deductive context, people reason about MP more naturally than other conditional schemata (Johnson-Laird & Byrne, 2002). This suggests that some participants may be engaging in a different kind of thinking when judging MP in comparison to the other schemata. Perhaps more abstract thinking leads to rejection of MP based on the ability to think of specific counterexamples without regard to their probability, in which case the number of disablers may be more important than W_c. This is consistent with work by Verschueren, Schaeken and d'Ydewalle (2005) showing two processes in causal conditional reasoning: A relatively quicker intuitive process that arrives at judgments that are highly correlated with conditional probability and a relatively slower, analytic process that correlates with number of alternatives or disablers.

5.3.3 Conclusions

With respect to the goals of the dissertation, the key takeaway from this chapter is what the reaction times imply about the processes underlying prediction and diagnosis. One

important finding is that across all conditions, diagnosis was slower than prediction. This is consistent with the idea that diagnosis is more difficult or less natural (Medin et al. 2003; Tversky & Kahneman, 1982). Diagnosis differs from prediction in that people consider the strength of alternative causes and prior probability, neither of which is used in prediction.

Diagnosis was faster with few alternatives. One explanation for this is that retrieval time increases with the number of alternative causes retrieved from memory. Speaking in favor of this hypothesis is that diagnostic judgments are higher under time pressure (Dougherty & Hunter, 2003b), suggesting that time pressure reduces the number of alternatives people can retrieve. Reaction time for predictions did not vary with number of alternative causes, providing more evidence that people neglect alternative causes when making predictions.

Reaction time for predictions did not vary with number of disablers but did vary with causal power. Prediction was slower when causal power was low. Differences in predictive reaction time could reflect more time spent retrieving disablers or a more difficult evaluation of conditional likelihood when the causal model includes a greater number of disablers and hence has a lower causal power. The lack of correlation between causal power and number of disablers suggests that causal power judgments are made by considering only the most important disablers, and the good fit between causal power and predictive likelihood suggests that the same is true of prediction. This idea squares with Thomas et al.'s (2008) claim that people tend to generate hypotheses that are highest in "*a priori* probability" (e.g. Dougherty & Hunter, 2003a). Because of this, number of disablers is not always a good proxy for causal power or good predictor of likelihood.

Beyond the reaction times, a secondary conclusion is that the work in this chapter corroborates the basic findings of Chapter 2. The full diagnostic model provided a close fit to diagnostic likelihood judgments, while W_c provided a good fit to predictive judgments. This provides more evidence that people use causal models to make inferences, but that their causal models for predictive reasoning do not include alternative causes.

Finally, the results suggest a fairly broad scope for the causal model conjecture. AC judgments were fit well by the diagnostic model suggesting that people were basing AC judgments on conditional likelihood derived from a reasonable causal model, as when asked for diagnostic likelihood. The caveat to this is that MP judgments seem not to be based as closely on causal power as predictive likelihood judgments. This may be due to participants using a mixture of strategies for MP. Of course, it's important not to jump to firm conclusions on the basis of so few examples (the poor fit to MP was primarily driven by 4 data points). Future work should aim to corroborate the differences in ratings and reaction times for MP versus predictive likelihood with a larger number of well-controlled items.

6. Development of Predictive and Diagnostic Reasoning[12]

6.1 Introduction

The work described in Chapters 2-5 suggests that the ease of predictive reasoning stems from people's tendency to focus on a particular cause and think forward from that cause to estimate the likelihood of the effect. Conversely, diagnosis requires a search through memory for other relevant causes and a comparison process between those causes and the one under consideration. These processes are demanding and time consuming. It stands to reason that young children should be capable of prediction, because the predictive processes allows them to focus on a single causal mechanism. Diagnosis should be slower to develop because the requisite processes are more demanding, necessitating the consideration of multiple hypothesis (cf. Beck et al., 2006).

Little work has compared children's predictive and diagnostic reasoning abilities. Bindra, Clarke and Shultz (1980) embedded a predictive/diagnostic manipulation within a larger experiment that was primarily aimed at comparing logical and causal reasoning. The causal task was to predict the appearance of a light based on the position of two switches and to diagnose the position of the switches based on the light. The task was quite complex as the functional relation between switches and light was also varied between conditions. Performance on predictive questions was superior to the diagnostic questions in general, supporting the idea that prediction develops earlier. However, the complexity of the design makes it difficult to tell whether the poor performance for diagnosis was due to the difficulty of reasoning about a variety of functional relations between switches and lights, or due to difficulty in considering multiple hypotheses.

In a more recent study Hong et al. (2004) assessed predictive and diagnostic reasoning using a variant of the "ramp task" (Frye, Zelazo & Palfai, 1995). The apparatus was composed of two tubes and a marble could be placed in either. Depending on condition, the tubes were set up such that the marble emerged either from the same tube it was placed in, or from the opposite tube. Based on training, children were fairly adept at predicting where the marble would emerge based on where it was put in, but were somewhat worse at choosing where to put it in to make it come out at a particular location. The latter is like the diagnostic tasks described throughout the dissertation in that it requires thinking backward from effect to cause but it also different in that the inference is hypothetical. Rather than observing the effect and diagnosing the cause, participants were asked to choose a tube to accomplish a goal (i.e. getting the ball to emerge from a particular location). Another difference is that in this task, the causal structure is such that there is only one possible cause for a given effect. The difficulty children experienced in diagnosis seems to be due to difficulty figuring out which of the two hypotheses applies to the given effect as opposed to an inability to consider both as possibilities.

In sum, both of these studies suggest that prediction develops earlier than diagnosis, but neither methodology is sufficient to assess the hypothesis that the late development of diagnosis reflects the difficulty of considering multiple hypotheses. Experiment 1 was designed to test a novel methodology for assessing predictive and diagnostic reasoning that does accomplish this objective.

6.2 Experiment 1

I used the 'blicket director' paradigm (Gopnik & Sobel, 2000) but adopted it so that participants were asked to make predictive and diagnostic inferences. In this paradigm, blocks are placed on a machine, which lights up and plays music in the presence of certain blocks. Children learn which blocks are effective and which are not and can subsequently be tested on their ability to transfer that knowledge to novel inferences. The paradigm therefore serves as a general tool for studying causal reasoning.

In an analogous fashion to the experiments in Chapter 2, I varied the number of alternative causes by teaching the children that either one or two of three blocks was effective. I obtained predictive inferences by asking whether a block would activate the machine it if it were placed on it. I obtained diagnostic inferences by occluding the machine and then activating it so that the child could not see which block was on it, and then asking the child which block had been used. To the best of my knowledge this methodology has not been used before, thus Experiment 1 served as a test of the methodology. I chose the simplest manipulation I could think of and predicted good performance. Critically, none of the tests in the experiment required considering multiple hypotheses since and I therefore did not expect any detriment to diagnostic performance.

6.2.1 Methods

Participants

9 three-year-olds (mean age = 39.6 months, 5 male) and 10 four-year-olds (mean age = 54.2 months, 6 male) were recruited from birth records.

Materials

The 'blicket detector' described by Gopnik and Sobel (2000) was used. The detector is a square box with a lucite top that depresses when an object is placed on it. The experimenter controls the detector's behavior via foot pedal. When the foot pedal is depressed the detector flashes red and plays music.

Two groups of three blocks each were also used. The blocks within each group were the same shape but different colors. All six blocks were different colors. When a block was placed on the detector it either 'activated' it or failed to do so. When demonstrating an active block, the experimenter surreptitiously depressed the foot pedal while simultaneously placing the block on the detector, causing the detector to flash and play music. She released the foot pedal upon removing the block. Thus it appeared to the child as if the block caused the detector to activate. For demonstrations where the block was ineffective, the experimenter placed the block on the detector but it did not flash or play music. Additionally, a piece of cardboard (approximately 2' x 2') was used to occlude the detector from the child when demonstrating a diagnostic event.

Design and Procedure

There were two key independent variables, direction of inference (predictive or diagnostic) and number of alternative causes (one-cause condition or two-cause condition), which were manipulated within-participant. Participants therefore provided responses to four kinds of trials.

Children were tested by one of two unfamiliar experimenters (1 male, 1 female). They sat facing the experimenter across a table. The experimenter placed the detector on the table and introduced the child to it by saying, "This is my machine. Some things make

it go and some things don't". The experimenter then placed three blocks on the table in front of the machine, all of the same shape but different colors.

Each of the blocks was placed on the machine one at a time for approximately three seconds starting with the block on the experimenter's left. In the one-cause condition one of the blocks activated the machine and two failed to do so. In the Two-cause condition two blocks activated the machine and one failed to do so. The spatial location of effective and ineffective blocks was randomized. After demonstrating each block, the experimenter repeated the demonstration. The child therefore either saw each block activate the machine twice or fail to do so twice.

After the training, the child performed a diagnostic and a predictive test trial. The order of these trials was counterbalanced such that the predictive was first half the time and diagnostic was first in the other half. For predictive trials the experimenter pointed to each of the three blocks in turn and asked the child "If I put this block on the machine, will it make the machine go?' After obtaining a verbal yes/no response the experimenter moved on and asked about the next block until the child had made a prediction about each block.

On diagnostic trials the experimenter placed the occluder in front of the detector and blocks so that they were not visible to the child. The experimenter then said "I'm going to put one of the blocks on the machine, so pay attention." The experimenter then placed a block on the machine and activated the detector so that it played music for approximately three seconds. The child could not see this, but could hear the music playing. The experimenter then removed the occluder and asked the child "Which of the

blocks did I put on the machine?" The trial was completed when the child pointed at one of the three blocks.

After completing both the predictive and diagnostic trials the experimenter moved on to either the one-cause or the two-cause condition (depending on which had been completed first) by putting away the first set of blocks and bringing out a new set. The order of the one-cause and two-cause conditions was counterbalanced.

6.2.2 Results and Discussion

As predicted, performance on both predictive and diagnostic tasks was close to ceiling. Predictive trials were coded as errors if the child made a mistake on any of the three blocks (i.e. predicted an effective block would fail to activate the detector or predicted an ineffective block would activate the detector). Two three-year-olds and one four-year-old evidenced a 'yes bias' responding 'yes' to all six predictive questions across both the one and two-cause conditions. One three-year-old responded 'yes' to all the blocks in the one-cause condition but answered correctly in the two-cause condition. Otherwise, there were no errors on predictive questions. Diagnostic trials were coded as errors if the child chose an ineffective block (one possibility in the two-cause condition, two possibilities in the one-cause condition). There were errors on just 2 of the 38 trials, which was superior to chance performance (41.7%), $z=14.2$, $p<0.00001$.[13]

The results of Experiment 1 suggest that three and four-year-olds were able to make successful predictions and diagnoses in a simple task. However, the method does not provide insight into the key feature of diagnostic reasoning, the ability to consider multiple hypotheses. In both the one and two-cause conditions, children could have succeeded by considering a single cause. In the two-cause case there was no way to know

whether they appreciated that there was an alternative explanation to the one that they had chosen or whether they had simply resolved with certainty on a single cause.

6.3 Experiment 2

The purpose of Experiment 2 was to test whether three and four year-olds could consider multiple hypotheses in diagnostic reasoning. To test this I used a similar method to Experiment 1 but introduced a fourth, novel block; participants did not know whether it was effective or not. In diagnostic test trials, participants made a judgment as before but were then told that their first guess was wrong and were asked to choose another block. If they are able to consider alternative causes then in the two-cause case they should shift their guess to the other effective block or to the novel block. In the one-cause case they should choose the novel block.

I predicted that diagnostic performance would be poorer than Experiment 1 due to an inability to consider multiple hypotheses. Conversely, I expected predictive performance to be good since as in Experiment 1, prediction did not require considering alternatives. This pattern would also suggest that diagnostic failures were not due to errors of memory because success on the predictive task implies that participants were able to remember which blocks were effective and which not.

I also expected a developmental difference. Four-year-olds are capable of a variety of sophisticated causal inferences that elude three-year olds (e.g. Sobel, Tenenbaum & Gopnik, 2004) and I therefore suspected that four-year-olds would be somewhat better at the diagnostic task. In other words I hypothesized that the ability to consider alternative causes in diagnostic reasoning begins to develop around age four.

6.3.1 Methods

Participants

15 three-year-olds (mean age = 40.0 months, 7 male) and 18 four-year-olds (mean age = 54.7 months, 11 female) were recruited from birth records.

Materials

Materials were identical to Experiment 1 except that there were now 4 groups of 4 blocks each. The blocks within each group were the same shape but different colors. Each group was a different shape and all 16 blocks were different colors.

Design and Procedure

The design was the same as Experiment 1 except that the predictive and diagnostic trials were separated into separate conditions with different training phases. The predictive and diagnostic trials for a given number of causes always occurred one after the other. The order of the one-cause versus two-cause conditions was counterbalanced, as was the order of predictive and diagnostic trials.

The procedure for the training phases was similar to Experiment 1 except that after demonstrating the three blocks, the experimenter brought out a fourth block and placed it on the table to the right of the other blocks. The experimenter then proceeded with a predictive or diagnostic trial. Predictive trials were the same as Experiment 1 except that the experimenter also asked a predictive question about the novel block. Diagnostic trials were also similar to Experiment 1 except that after the child made a guess the experimenter removed the chosen block from the table and responded “That’s a good guess, but actually it’s not right. That’s not the one I put on the machine. Can you

tell me which one I did put on the machine?" The trial concluded when the child chose another block.

6.3.2 Results

Overall Performance on Prediction and Diagnosis

As in Experiment 1, predictive trials were coded as errors if any of the predictive questions in the trial were answered incorrectly. Either 'yes' or 'no' was considered correct for the novel block. Again, some participants evidenced 'yes bias'. Two three-year-olds and 1 four-year-old answered 'yes' to all eight predictive questions. Unless noted, these responses were removed from the dataset for subsequent analyses. Diagnostic trials were coded as errors if the child guessed an ineffective block in either the first or second guess or both. Chance performance differed across the one-cause and two-cause conditions due to the differing number of ineffective blocks. Table 6.1 shows overall performance for prediction and diagnosis.

Table 6.1: Percentage of trials without errors for prediction and diagnosis. Chance performance is in parentheses.

	Prediction	*Diagnosis*
3-year-olds	76.9% (12.5%)	40.0% (33.3%)
4-year-olds	91.2% (12.5%)	50.0% (33.3%)

Note: Chance performance was determined by computing the probability of at least one error in a trial assuming uniform choice over options.

A Z-test comparing proportions revealed that performance on predictive trials was superior to diagnostic trials overall, $Z=4.6$, $p<0.00001$. This was true even when 'yes bias' responses were included, $z=3.8$, $p<0.001$. The difference was also significant for three year olds ($Z=2.8$,$p<0.01$) and four year olds ($z=3.8$, $p<0.001$) separately.

Performance on predictive trials was much better than chance ($z=6.1$, $p<0.00001$). Diagnostic performance was not better than chance ($z=1.5$, $p>0.05$). Three-year-olds and four-year-olds did not differ from each other on predictive ($z=0.8$, $p>0.05$) or diagnostic trials ($z=1.7$, $p>0.05$).

One-cause and Two-cause Conditions

To assess the effect of number of alternative causes, I looked at the data for each condition separately. Table 6.2 shows the percentage of correct diagnostic trials by condition for each age group. Overall, in neither condition was performance better than chance. However, four-year-olds were above chance in the one-cause condition, $z=1.8$, $p<0.05$ and were better than the 3-year-olds, $Z=3.6$, $p<0.001$. Thus, while performance overall was quite poor for both groups this provides some evidence that some 4 year olds were able to accomplish the diagnostic inference in the one-cause case.

Table 6.2: Percentage of diagnostic trials without errors by condition and age. Chance performance is in parentheses.

	Two-causes	*One-cause*
3-year-olds	66.7% (50.0%)	13.3% (16.7%)
4-year-olds	55.6% (50.0%)	44.4% (16.7%)

Note: Chance performance was determined by computing the probability of at least one error in a trial assuming uniform choice over options.

Predictive performance did not vary as a function of number of causes. There were more errors in the two-cause condition (20% of trials) than the one-cause condition (10% of trials) but this difference was not significant, $z<1$, *ns*.

To try to understand the differential diagnostic performance by 4-year-olds in the one-cause condition, I looked at whether the diagnostic errors in each condition were

made on the first guess or the second guess. A suggestive difference between conditions emerged. In the two-cause condition, participants were more likely to make an error on the first guess (9 of the 13 errors were on the first guess) while in the one-cause condition errors on the second guess were more common (18 of 23 errors were on the second guess). This difference was significant, $z=5.6$, $p<0.00001$. I speculate about this difference in the discussion.

Novel Block Choice

One might ask whether children's tendency to make diagnostic errors reflects a failure to understand that the novel block could activate the machine. At least two findings speak against that interpretation. First, participants often chose the novel block in diagnostic trials suggesting that they did understand that it was a possible cause. Table 6.3 shows the distribution of block choices across diagnostic trials. (Note that in the two-cause condition the number of effective blocks is two and in the one-cause condition it is one). The prevalence of novel block choices makes it unlikely that diagnostic failures resulted from a misunderstanding of the instructions. Participants treated the novel block similarly to the other blocks. Second, in predictive trials participants tended to predict that the novel block would activate the detector and this did not vary across the number of causes (26 of 33 trials in the two-cause condition and 28 of 33 in one-cause condition). This suggests that if anything, there was a bias to interpret the novel block as effective in the absence of definitive evidence.

Table 6.3: Distribution of Responses on Diagnostic Trials

	Two-Cause Condition	One-Cause Condition
Effective Block(s)	40	23
Ineffective Block(s)	13	25
Novel Block	13	18

6.3.3 Discussion

I made two predictions about Experiment 2. I predicted that diagnostic performance would be poor relative to Experiment 1 while predictive performance would remain strong. This prediction was borne out. Diagnostic performance was consistent with chance while prediction was near ceiling. The second prediction of a developmental difference was partially supported. Four-year-olds while still performing poorly overall were better than chance and better than the three-year-olds in the one-cause condition. Three-year-olds were at chance in both conditions.

The predictive success makes it unlikely that diagnostic errors were due to simple memory failures. Participants were able to remember which blocks were effective and which ineffective, but they had trouble using that knowledge to their advantage in diagnostic inference. Diagnostic failures also were not due to misunderstanding the role of the novel block. Participants treated it much like the other blocks.

An interesting trend emerged by looking at whether diagnostic errors occurred on the first or second guess. In the two-cause condition errors tended to occur on the first guess while in the one-cause Condition they tended to occur on the second guess. A possible explanation for this is difficulty in considering multiple hypotheses. On this explanation, in the one-cause condition participants tend to focus on the one effective cause. This allowed them to succeed in Experiment 1, and to do well on their first guess

in Experiment 2. But when they are told that their guess is wrong, they have trouble revising their opinion and adopting a new hypothesis. The fact that 4-year-olds were somewhat better in this condition suggests that this capability may be just beginning to emerge.

In the two-cause condition, there were fewer errors (due to the presence of only a single ineffective block) and therefore conclusions are harder to draw. Still the prevalence of errors on the first guess differs from the results of Experiment 1 suggesting that the presence of the novel block is interfering with the children's ability to identify even a single good explanation. As discussed above, it does not seem to be the case that the extra information is impinging on their ability to remember which blocks are effective, but rather is disrupting their ability to converge on a hypothesis when making the diagnostic inference.

6.4 General Discussion

Two experiments tested 3 and 4-year-olds' ability to make predictive and diagnostic inferences. I predicted that the ability to consider multiple hypotheses begins to emerge at age 4 but that prediction emerges earlier. In Experiment 1 the children could succeed without considering multiple hypotheses and performance was near ceiling in both prediction and diagnosis. In Experiment 2 I forced children to consider a different hypothesis in diagnosis by telling them their first guess was wrong and asking them to choose another block. Overall, diagnostic performance was poor, though 4 year-olds were somewhat better in the 1-cause condition suggesting that 4-year-olds are just beginning to be able to consider multiple hypotheses. As expected, predictive performance remained strong.

Given the substantial difference in methodology of Experiment 1 versus Experiment 2, one might ask whether a more minimal change might provide useful evidence for or against that hypothesis that the failures in Experiment 2 are due to an inability to consider multiple hypotheses. Such an experiment is currently in the data collection stage. The experiment replicates the two-cause condition of Experiment 1, but includes the innovation from Experiment 2 of removing the child's first guess and requesting a second guess. In the absence of a novel block, will 3 and 4-year olds still find it difficult to shift to a new hypothesis in diagnostic trials?

The results suggest that the development of causal reasoning begins with the ability to represent individual causal relations and to use knowledge about that relation to predict the effect when the cause is present. This kind of local focus requires minimal processing and may be quite a good strategy to fulfill needs in early child hood. It even supports a limited kind of diagnostic reasoning, the ability to reason backward to a particular cause, as in the one-cause condition of Experiment 1. But this kind of reasoning does not bear the hallmark of full diagnostic inference: the consideration of multiple hypotheses and the understanding that more than one cause is a potential explanation for a given outcome. This capability is likely more useful when a child begins to reason about more complex causal systems and has more nuanced goals. It may only convey a benefit later in development.

7. Conclusions

7.1 Summary of Results

I have reported 5 lines of work exploring predictive and diagnostic reasoning. In the first line, I proposed a normative causal model theory of transmission arguments based on the noisy-or common effect model. I derived equations for predictive and diagnostic conditional probability based on the underlying parameters of the causal scenario, prior probability, causal power and strength of alternatives. In the experiments, the model was found to do a fairly good job of predicting mean judgments. This suggested that people used causal models to make likelihood judgments. They distinguished questions based on causal direction and their judgments were sensitive to many of the right things. There was one systematic violation of the normative model; Prediction was insensitive to alternative strength. People only considered the causal strength of the cause mentioned in the question. This neglect of alternative causes was shown not to be due to people misinterpreting the questions. When mentioned explicitly in the question, alternative causes were considered. Moreover, the pattern of results was identical when questions were phrased in unambiguous frequency language. This line of work cast doubt on the causal asymmetry conjecture. There was no support for the idea that predictions are judged higher than diagnoses all else being equal. There was however evidence of the opposite. Predictions were too low due to the neglect of alternative causes. Diagnoses were unbiased with respect to people's underlying beliefs.

In the second line of work I took a more direct approach to assessing the consideration of alternative causes in prediction and diagnoses. I compared judgments of 'full' and 'no-alternative' conditionals. No-alternative conditionals ruled out causes other

than the one mentioned in the questions. The results of three experiments were the same: participants neglected alternative causes in prediction but treated them appropriately in diagnosis. This was established with mental health professionals reasoning about a case (Experiment 1); and undergraduates reasoning about goals and means (Experiment 2) and property transmissions (Experiment 3).

In the third line of work I explored the magnitude of errors people make due to the neglect of alternative causes. The potential for error is great when the focal cause is weak and the strength of alternative is high. In an extreme case, neglect of alternative leads to positive evidence reducing belief, a phenomenon I refer to as the weak evidence effect. The effect was established by comparing judgments of marginal likelihood to judgments of condition likelihood given a weak cause about 4 public policy issues. Conditionals were judged lower. In a separate condition the causes were judged probability raising. The weak evidence effect was corroborated with questions about everyday causal scenarios. I also obtained judgments of causal power in this experiment and found them to be lower than the conditional judgments, suggesting that people did sometimes think about alternative causes in prediction, but not sufficiently.

In the fourth line of work I obtained reaction times for predictive and diagnostic likelihood judgments about a variety of causal scenarios drawn from Cummins (1995). The scenarios varied in the number of disabling conditions and in the number of alternative causes. In all conditions, prediction was faster than diagnosis supporting the idea that prediction is more natural or easier. One reason for this difference is the retrieval and evaluation of alternative causes in diagnosis but not in prediction. Supporting this, diagnosis was slower for items with many alternative causes. Prediction

did not vary with number of alternative causes, providing more evidence of neglect of alternatives. Prediction did vary with causal power. It did not vary with number of disablers however. This may because only high probability or common disablers are used for prediction. On this explanation, the retrieval and/or evaluation of these high probability disabler leads to slower predictive judgments. Another outcome of this line of work was to show that some deductive reasoning tasks also draw on causal models. This was revealed by the good fits of the normative model to Cummins' affirming the consequent acceptability ratings.

In the fifth line of work I tested the hypothesis that diagnosis develops later than prediction because of the computational demands imposed by the retrieval and evaluation of multiple hypotheses. Two experiments used a variation on the blicket detector paradigm; Children were asked to predict whether a particular block would activate the machine or to diagnosis which block had the activated the machine when it was occluded. Overall, performance on prediction was near ceiling. Performance on diagnosis was poor, and consistent with chance. However, 4-year-olds appeared to be just beginning to develop the ability to consider multiple hypotheses based on their slightly above chance performance in the 1-cause condition.

7.2 Speculation on Underlying Mechanisms

In Chapter 6, I suggested that children's proficiency for causal reasoning begins with the ability to reason about individual causal mechanisms. I speculate that this ability is basic to how causal knowledge is represented in adults as well. On this account, the basic representational unit is the individual causal mechanism. A causal mechanism includes a causal path from cause to effect and normal or prototypical enablers and disablers. The

components of a mechanism 'hang together' in memory. This makes prediction relatively low effort. When people think about a mechanism they automatically think about the elements necessary for deriving a reasonable approximation to causal power, which in turn allows them to predict the effect by simulating the mechanism (i.e. running it forward), as discussed in Chapter 3.

People prefer to think about as few mechanisms at a time as possible and therefore ignore alternative mechanisms when making predictions. As discussed in Chapters 1 and 2 this tendency is an instantiation of a more general phenomenon to ignore relevant information across a wide variety of judgment and decision-making tasks. Violations of extensionality arise because judgment depends on which causal mechanism happens to be the focus of attention. When provided with a candidate mechanism -- as when I asked people conditional likelihood questions – this candidate is the focus of attention. Likelihood judgments can therefore be pushed around by conditioning on particular causes.

When asked for the marginal likelihood, people retrieve the most important cause or causes to make an estimate (cf. Dougherty & Hunter, 2003a). Non-monoticity can occur because drawing attention to a weak causal mechanism leads people to focus on it to the exclusion of other causes. This is similar to Sloman et al.'s (2004) 'narrow interpretation conjecture.' Marginal effect likelihoods are based on the most typical or important causes.

Diagnostic judgment differs from predictive in that alternative causal mechanisms cannot be ignored while making reasonable judgments because diagnosis is inherently comparative. This was shown in the normative analysis in Chapter 2. While alternative

causes are relevant to predictions, they make an independent contribution to the effect. Ignoring alternatives leads to error, but the resultant predictive judgment is at least correlated with the normative likelihood. Due to the asymmetry of causal relations, the same is not true for diagnostic inference. In normative diagnostic likelihood there is dependence between the likelihood of independent alternative causes. If one cause is a better explanation for an effect, an alternative is necessarily a poorer explanation. Ignoring alternatives leads to incoherence. People therefore have no choice but to do the hard work of thinking of alternatives and keeping multiple mechanisms in mind to make a judgment.

Evidently, people do eventually learn how to make reasonably good diagnostic inferences, at the cost of a cognitive process that is relatively time consuming and demanding. One possibility for how this works is that diagnostic inference is scaffolded onto the basic ability to reason about individual causal mechanisms. On this explanation, when faced with a diagnostic inference, people begin by representing the focal causal mechanism as in prediction, but then add detail to the representation that includes prior probability and alternative causal mechanisms. Alternatives come to mind in order of strength by their association with the effect. Presumably, participants do not compute the exact diagnostic equation. Likely, they estimate it as a simple function that compares the relative strength of the focal cause to the most important alternatives, in line with Support Theory. Future work should aim to articulate this heuristic in detail.

7.3 Final Thoughts

One conclusion of this work is that accurate descriptive models of intuitive likelihood assessment will have to accomplish two things: (a) account for people's sensitivity to the

normative principles that distinguish prediction and diagnosis and (b) model the cognitive processes that support these inferences to account for errors of judgment. Prevailing models do one or the other, but not both. Exploring and formalizing how people construct causal models from their conceptual knowledge thus offers a promising avenue for future research.

Notes

Chapter 2

1. The work reported in this chapter is currently under review (Fernbach, Darlow & Sloman 2009). Experiment 1 and aspects of the model were published in the 2009 Proceeding of the Cognitive Science Society (Fernbach & Darlow, 2009). The work was supported by NSF Award 0518147 to Steven Sloman and by a Brown University Fellowship and an APA Dissertation Research Award to Phil Fernbach. I thank Jonathan Bogard for help collecting data and David Over and Dinos Hadjichristidis for helpful discussions.

2. According to the PowerPC model of causal learning, causal powers are inferred from contingency data on the assumption that causes contribute to effects independently (i.e., a according to a noisy-or model). My model captures inference rather than learning. Causal power is given and conditional likelihoods of causes and effects are inferred.

3. Some of the categories could be described as having part-whole relationships, but I still consider them transmission scenarios because the predicate applies to the part before it applies to the whole. Importantly, the predicates were such that the if the predicate applies to the part, it increases the probability of the predicate applying to the whole but does not make it necessary. Therefore, the causal structures of these items do not differ from the rest.

4. Because each participant did not make the same number of judgments for each dependent variable, each participant did not supply a sufficient number of judgments per condition to support an analysis by participants. I therefore collapsed over participants and used the category means for all of the analyses of Experiment 1.

5. Given the relatively high causal powers for the items in all the experiments in Chapter 2 a reasonable question is what happens when causal powers are lower. The evidence on this question is mixed. Chapter 4, Experiment 2 did show partial consideration of alternatives with weak causal powers. Ongoing work with Bob Rehder using artificial categories has shown no consideration of alternatives when causal powers are weak (equal to 0.4).

6. It should be noted that the claim that diagnostic reasoning is unbiased refers to consistency between parameter judgments and conditional likelihood judgments. Thus diagnostic reasoning could still be biased with respect to the true probabilities if both parameter judgments and likelihood judgments are biased in the same way. For instance, W_a was probed by asking for the likelihood of the effect in the absence of the cause. In that case, people may think of some alternatives but not all of them. If the same is true when making the diagnostic judgment, there would be consistency but bias with respect to the true probabilities. W_a judgments would be too low and judgments of D too high.

7. Methods were as follows: Participants were approached on the Brown University campus and participated voluntarily. They were asked either the predictive or diagnostic question verbally and the experimenter wrote down their responses. Responses were analyzed with an independent samples t-test.

Chapter 3

8. The work described in this chapter is taken from a manuscript currently in press (Fernbach, Darlow & Sloman, 2010a). The data from Experiment 1 appeared in a poster presented at the Simches Symposium, Boston, MA, 2008 (Romeo et al., 2008). I thank

Jonathan Bogard for collecting data and Leonel Garcia-Marques, Ju-Hwa Park, and John Santini for discussions of the work.

9. The pattern of results in Experiment 1 is slightly different in the relatively lower value for the no-alternative diagnostic conditional. I suspect that this is due to the question specifying that 'A complete diagnostic workup revealed that she has not been diagnosed with *any other medical or psychiatric disorder* that would cause lethargy" (emphasis added). The vignette ruled out other medical or psychiatric causes of lethargy but not all possible causes (i.e. causes that are not medical or psychiatric), which could have led to relatively lower values for the no-alternative diagnostic conditional.

Chapter 4

10. The work described in this chapter is taken from a manuscript that is currently in preparation (Fernbach, Darlow & Sloman, 2010b). The work was supported by a Brown University Fellowship and an APA Dissertation Research Award to myself.

Chapter 5

11. The work described in this chapter is taken from a manuscript to be published in the 2010 Proceedings of the Cognitive Science Society (Fernbach & Darlow, 2010). The work was supported by a Brown University Fellowship and an APA Dissertation Research Award to the first author. I thank Steve Sloman, David Over and Dinos Hadjichristidis for helpful discussion and am especially grateful to Denise Cummins for digging up her data from 1995.

Chapter 6

12. Dave Sobel provided valuable advice in designing and executing this research. Dave Buchanan and Deanna Simeone collected the data and Dave Buchanan provided helpful

advice and discussions. The members of the Causality and Mind Lab also gave useful feedback.

13. All Z-tests in this chapter use two tails when comparing conditions to each other and one-tail when comparing to chance.

References

Ahn, W. & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, 31, 82-123.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology. General*, 136(4), 569-76.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Beck, S. P., Robinson, E. J., Carrol D. J. & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development,* 77 (2), 413-426.

Bindra, D., Clarke, K. A. & Shultz, T. R. (1980). Understanding predictive relations of necessity and sufficiency in formally equivalent "causal" and "logical" problems. *Journal of Experimental Psychology: General,* 109 (4), 422-443.

Blok, S. V., Medin D. L. & Osherson, D. N. (2003). Induction as conditional probability judgment. *Memory and Cognition*, 35 (6), 1353-1364.

Bornstein, B.H., & Emler, A.C. (2001). Rationality in medical decision making: A review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice*, 7, 97–107.

Chater, N., & Oaksford, M. (Eds.) (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford: Oxford University Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104, 367–405.

Cummins, D. D (1995) Naïve theories and causal deduction. *Memory and Cognition*, 23 (5), 646-658.

Cummins, D. D., Lubart, T., Alksnis, O. and Rist, R. (1991) Conditional reasoning and causation. *Memory and Cognition*, 19 (3), 274-282.

Dawes, R. M. (2001) *Everyday Irrationality: How pseudoscientists, lunatics, and the rest of us fail to think rationally*. Boulder, CO: Westview Press.

De Neys, W., Schaeken, W. & D'ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory and Cognition*, 30 (6), 908-920.

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D. & Mynatt, C. R. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Memory & Cognition*, 24 (5), 644-654.

Doherty, M. E., Mynatt, C. R., Tweeney, R. D. & Schiavo, M. D. (1979). On pseudodiagnosticity. *Acta Psychologica*, 43, 111-121.

Dougherty, M.R.P., Gettys, C.F., & Thomas, R.P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, 70, 135–148.

Dougherty, M. R. P., Gettys, C. F. & Ogden, E. E. (1999). Minerva-DM: A memory process model for judgments of likelihood. *Psychological Review*, 106 (1), 180-209.

Dougherty, M. R. P. & Hunter, J. E. (2003a). Hypothesis generation, probability judgment and individual differences in working memory capacity. *Acta Psychologica*, 113, 263-282.

Dougherty, M. R. P. & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory and Cognition*, 31, 968-982.

Elstein, A. S. Shulman, L. S. & Sprafka S. A. (1978). Medical problem solving: An analysis of clinical reasoning. Cambridge, MA: Harvard University Press.

Evans, J. ST. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. London: Taylor & Francis.

Evans, J. St. B. T., Over, D. E. & Handley, S. J. (2003). A theory of hypothetical thinking. In D. Hardman & L. Maachi (Eds.). *The psychology of reasoning and decision making*. (pp. 3-21). Chichester: Wiley.

Fernbach, P. M. & Darlow, A. (2009). Causal asymmetry in inductive judgments. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

Fernbach, P. M. & Darlow, A. (2010). Causal conditional reasoning and conditional likelihood. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

Fernbach, P. M., Darlow, A. & Sloman, S. A. (2010a). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*. In Press.

Fernbach, P. M., Darlow, A. & Sloman, S. A. (2010b). *When good evidence goes bad: The weak evidence effect in predictive judgment*. Manuscript submitted for publication.

Fernbach, P.M., Darlow, A. & Sloman, S.A. (2009). *Asymmetries in predictive and diagnostic reasoning*. Manuscript submitted for publication.

Fernbach, P. M. & Darlow, A. (2009). Causal asymmetry in inductive judgments. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

Fox, C. R. & Levav, J. (2004). Partition-edit-count: Naïve extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, 133(4), 626-642.

Fox, C. R. & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, 14(3), 195-200.

Fischhoff, B., Slovic, P. & Lichtenstein (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human, Perception, and Performance*, 4, 330–344.

Frye, D., Zelazo, P. D. & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, 10, 483-527.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102 (4). 684-704.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: Bradford Books.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. Psychological Review, 111 (1), 3-32.

Gopnik, A. & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development* 71, 1205-1222.

Hadjichristidis, C., Sloman, S. A., & Over, D. E. (2009). *Categorical induction from uncertain premises: Jeffrey's (doesn't) rule*. Manuscript submitted for publication.

Hagmayer, Y., & Waldmann, M.R. (2000). Simulating causal models: The way to structural sensitivity. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 214–219). Mahwah, NJ: Erlbaum.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition*: Oxford University Press.

Hertwig, R. & Gigerenzer, G. (1999). The `conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12, 275-305.

Hogarth, R. M. & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1-55.

Hong, L., Chijun, Z., Xuemei, G., Shan, G. & Chongde, L. (2004). The influence of complexity and reasoning direction on children's causal reasoning. *Cognitive Development*, 20 (1), 87-101.

Johnson-Laird, P. N. & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109 (4), 646-678.

Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M. S. & Caverni, J. (1999), Naïve probability: A mental model theory of extensional reasoning. *Psychological Review*, 106(1), 62-88.

Josephson, J.R., & Josephson, S.G. (1994). *Abductive Inference: Computation, philosophy, technology*. New York: Cambridge University Press.

Kahneman, D. & Tversky, A. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90 (4), 293-315.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.

Kelley, H. H. (1972). Causal schemata and the attribution process. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. S. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151-174). Morristown, NJ: General Learning Press.

Kemp, C., & Tenenbaum, J. (2003). Theory-based induction. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1), 20-58.

Keysar, B., Lin, S. & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89, 25-41.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211– 228.

Koehler, D. J., White, C. M. & Grondin, R. (2003). An evidential support accumulation model of subjective probability. *Cognitive Psychology*, 43, 152-197.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37, 2098-2109.

Lord, C.G., Lepper, M.R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, 47, 1231–1243.

McKenzie, C. R. M., Lee, S. M. & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making*, 15, 1-18.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review*, 10 (3), 517-532.

Nozick, R. (1993). *The nature of rationality*. Princeton. Princeton University Press.

Pearl, J. (2000). Causality. Cambridge: Cambridge University Press.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Science*, 5, 349 – 357.

Oaksford, M., & Chater, N. (2003). Conditional probability and the cognitive science of conditional reasoning. *Mind and Language*, 18 (4), 359 – 379.

Oaksford, M., Chater, N. & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning Memory and Cognition*, 26, 883-889.

Osherson, D. M., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, 97, 185-200.

Over, D., Hadjichristidis, C., Evans, J. St BT. Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, 54, 62-97.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.

Peirce, C.S. (1931).. In C. Hartshorn & P. Weiss (Eds.), *Collected papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press.

Pitz, G.F., Downing, L., & Reinhold, H. (1967). Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology*, 21, 381–393.

Quinn, S. & Markovitz, H. (1998). Conditional reasoning, causality and the structure of semantic memory: Strength of association as a predictive factor for content effects. *Cognition*, 68, B93-B101.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1141-59.

Rehder, B. (2006). When similarity and causality compete in category-based property induction. *Memory & Cognition* 34, 3-16.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science* 33, 301-344.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 14 (6), 665-681.

Romeo, S., Sutton-Skinner, K., Petersen, T., Baer, L., Huffman, J., Birnbaum, R., & Sloman, S.A. (2008). *Clinical decision making biases in a group of mental health providers*. Poster presented at the Simches Symposium, Boston, MA.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 736-753.

Rottenstreich Y. & Tversky A. (1997). Unpacking, repacking and anchoring: Advances in support theory. *Psychological Review*, 104, 406-415.

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. *Advances in Neural Information Processes* 15.

Shafto, P., Kemp, C., Baraff Bonawitz, E., Coley, J. D. & Tenenbaum, J. B. (in press). Inductive reasoning about causally transmitted properties. *Cognition*.

Shah, J.Y., Friedman, R., & Kruglanski, A.W. (2002). Forgetting all else: On the antecedents and consequences of goal shielding. *Journal of Personality and Social Psychology*, 83, 1261–1280.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25, 231-280.

Sloman, S. A. (2005). *Causal models; how people think about the world and its alternatives*. New York: Oxford University Press.

Sloman, S.A., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 573-582.

Sloman, S. A. & Lagnado, D. A. (2005). Do we 'do'? *Cognitive Science* 29 (1), 5-39.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition*, 49, 67-96.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303-333.

Spirtes, P., Glymour, C. & Scheines R. (1993). *Causation, prediction and search*. New York: Springer-Verlag.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629-640.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12, 435–502.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115 (1), 155-185.

Tversky, A., & Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology* 5, 207-232.

Tversky, A. & Kahneman, D. (1982). Causal schemata in judgements under uncertainty. In D. Kahneman, P. Slovic & A. Tversky (eds.), *Judgement under uncertainty: Heuristics and biases* (117-128). Cambridge: Cambridge University Press.

Tversky, A. & Koehler (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101 (4), 547-567.

Verschueren, N., Schaeken, W. & d'Ydewalle, G. (2005). A dual process specification of causal conditional reasoning. *Thinking & Reasoning*, 11 (3), 239-278.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology; Learning, Memory, and Cognition*. 26 (1), 53-76.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121, 222-236.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: a minimal rational model. In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian Cognitive Science* (pp. 453-484). Oxford: University Press.

Windschitl, P. D. & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, 75, 1411-1423.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Appendix A

Categories and Predicates Used in Chapter 2, Experiment 1

Cause Category	*Effect Category*	*Strong Alternatives Predicate*	*Weak Alternatives Predicate*
Mother	Newborn baby	Has dark skin	Is drug-addicted
Parents in NYC	Only child	Speak English as first language	Know child's birthday present
Coach	High school football team	Is motivated	Knows a complicated play
Commuter Train	Commuter	Is late	Passes through several stations
Machine for manufacturing lenses	Lens	Is defective	Has micrometer precision
Mayor of a major city	New Policy	Is unpopular	Is fiscally conservative
Hard disk	Computer	Is broken	Can't hold any more files
Wheels	Car	Fail Inspection	Are Moving Fast
Television manufacturers	Electronics Stores	Sold an above-average number of defective products in 2007	Introduced a TV based on a new standard in 2007
Oranges	Orange Smoothie	Are sweet	Are sour
Apple Slices used to make an apple pie	Apple Pie	Are sweet	Have seeds
Music at a party	Party	Is loud	Is good for dancing
Company from the NYSE	Senior Manager at the Company	Is doing well financially	Uses Blue Cross health insurance
Transfusion blood at African Hospital	Transfusion Patient	Has an infectious disease	Is anemic
Early Spring day in NYC	An apartment in NYC	Is warm	Is sunny
Engine of a 2005 Honda accord	2005 Honda Accord	Is not functioning properly	Smells of burnt oil
Northern Ash wood	Baseball bat made from the wood	Is dark in color	Is liable to split
Body of water	Stew made from fish that live in the body of water	Is salty	Is high in mercury
Oxygen tank	Scuba diver	Has insufficient oxygen	Has plenty of oxygen
Tap water	Ice cubes made from the tap water	Tastes bad	Contains fluoride

Stimuli from Chapter 2, Experiment 3

Cause Category	*Effect Category*	*Predicate*	*Alternative Cause*
Mother	Newborn baby	Has dark skin	A father with dark skin
Coach	High school football team	Is motivated	Accolades from family and friends
Hard disk	Computer	Is broken	Other parts of the computer being broken, like the power source or the motherboard
Television manufacturers	Electronics Stores	Sold an above-average number of defective products in 2007	Defective products that come from other sources, like computer manufacturers
Apple Slices used to make an apple pie	Apple Pie	Are sweet	Adding sugar
Music at a party	Party	Is loud	Loud conversations or other sources of loud noise
Early Spring day in NYC	An apartment in NYC	Is warm	A heater turned on
Northern Ash wood	Baseball bat made from the wood	Is dark in color	Dark paint or stain
Tap water	Ice cubes made from the tap water	Tastes bad	Being frozen next to something that has a strong odor

Appendix B

Additional goal schemata for Chapter 3, Experiment 2

Goal	*Means*
Learn French	Use Rosetta Stone Software
Weigh less in May than April	Exercise Hard
Complete a marathon	Train hard
Become a millionaire by age 40	Get a high-paying job out of college
Quit smoking	Use a nicotine patch
Get a good grade on a test	Study hard
Get into a serious relationship	Join a dating service
Get good at guitar	Take lessons

Additional categories and predicates used in Chapter 3, Experiment 3

Cause Category	*Effect Category*	*Strong Alternatives Predicate*	*Weak Alternatives Predicate*
Mother	Newborn baby	Has dark skin	Is drug-addicted
Coach	High school football team	Is motivated	Knows a complicated play
Hard disk	Computer	Is broken	Can't hold any more files
Television manufacturers	Electronics Stores	Sold an above-average number of defective products in 2007	Introduced a TV based on a new standard in 2007
Apple Slices used to make an apple pie	Apple Pie	Are sweet	Have seeds
Music at a party	Party	Is loud	Is good for dancing
Early Spring day in NYC	An apartment in NYC	Is warm	Is sunny
Engine of a 2005 Honda accord	2005 Honda Accord	Is noisy	Smells of burnt oil
Northern Ash wood	Baseball bat made from the wood	Is dark in color	Is liable to split
Tap water	Ice cube made from tap water	Tastes bad	Contains fluoride

Appendix C

Additional Stimuli For Chapter 4, Experiment 2

Theme	*Conditional*
Cell Phone	A woman is a 35 year old whose parents live in a different state. She loses her cell phone on April 1st. How likely is it she doesn't talk to her parents in April?
Beer Company	A beer company owns a leading light beer. The company increases the advertising budget for its light beer by 3 percent. How likely is it the beer gains market share in the next year?
Vineyard	A California vineyard specializes in French style wine. The vineyard imports top soil from France. How likely is it that the wine scores well in a blind taste test by French critics?
Probiotic Diet	A man is a 20 year old university student. He is on a probiotic diet. How likely is it he goes a year without the flu?
House Flipper	A house flipper is looking to sell a property he acquired one year ago. He repaints all of the bedrooms in the house. How likely is it he realizes at least a 2% profit when he sells?
College	A young man is applying to colleges and trying to improve his application. He volunteers for the big brother program. How likely is it he gets into a top 100 college?
Jacket	A young man is a healthy high school student. He goes out during a heavy rain without a jacket. How likely is it he gets a cold sometime this winter?
Gasoline	A woman has 2003 Honda. She uses the lowest grade of gasoline. How likely is it the car has mechanical problems in the next year?
Milk	A man buys a half-gallon of milk on Monday. The power goes out for 30 minutes on Tuesday. How likely is it the milk is spoiled a week from Wednesday?
Smoking	A 30-year-old woman wants to quit smoking. She goes to hypnosis sessions. How likely is it she no longer smokes in 1 year?
Baseball	A baseball player hit 20 homeruns in the 2009 season. After the season he used a computer program twice a week to train his visual acuity. How likely is it he hits more than 20 homeruns in the 2010 season?
Tourist	A tourist is taking a picture of the statue of liberty from the deck of the ferry. There is a breeze at the moment he takes the picture. How likely is it the photo comes out blurry?

Note: Alternative question forms (marginal, casual power and probability raising) were generated as in the example in the text.