Models of human causal learning: review, synthesis, generalization.

(A long argument for a short rule)

Colin S. Beam

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

John M. Miyamoto, Chair

John C. Palmer

Andrea Stocco

Program Authorized to Offer Degree:

Psychology

University of Washington

Abstract

Models of human causal learning: review, synthesis, generalization.
(A long argument for a short rule)

Chair of the Supervisory Committee:
Professor John M. Miyamoto
Department of Psychology

This dissertation is composed of three major components. The first reviews models of causal learning with special emphasis given to Bayesian approaches. The second component joins algorithmic and computational models by defining the free parameters of the former in terms of the theoretical constructs of the latter. Specifically, the weighted ΔP model can be naturally expressed as an estimator of Cheng's (1997) causal power. This allows for a computational analysis of weighted ΔP that results in a number of insights. The analysis suggests that previous formulations of preventive weighted ΔP have been misspecified. With the correct specification, weighted ΔP is shown to be the best fitting model when entered in to Perales and Shanks (2007) model competition study. The analysis also facilitates a novel derivation of a more general Rescorla-Wagner model that attains a causal power equilibrium. Weighted ΔP is non-Bayesian, though it shares some characteristics with Bayesian estimators. Like the posterior mean, weighted ΔP can be interpreted as a compromise between a prior expectation and sample information. As such, it is also a low variance estimator of causal power. In contrast to Bayesian models, weighted ΔP predicts deterministic strengths of 0 or 1 in certain experimental conditions. Experimental results support these predictions and lead to the discovery of a "deterministic bias" in causal judgments. This phenomenon is strongly inconsistent with Bayesian models, though it also poses problems for point prediction models more broadly. The third component of the dissertation proposes *capacity and response probability* (CARP), a latent variable framework for models of causal inference. Under CARP, causes are associated with latent capacities. Conjoined causes are assumed to combine additively in their capacities. A response function maps capacity to the judged probability of the effect. Different response functions imply different models of causal judgment. After establishing the framework, response functions are derived for the ΔP rule and causal power, and a number of additional applications of CARP are proposed.

Table of Contents

List of Figures

List of Tables

Introduction.

What are people doing when they form judgments of causal strength? How do they use evidence to form these judgments? In attempting to answer these questions, the field of causal learning has enjoyed a spirited debate between several distinct research approaches, with each giving different emphasis to the preceding two questions. Indeed, one can roughly categorize different research traditions in psychology based on the varying weight they give to "what" versus "how" questions.

Explanations in cognitive science often take the form of input-output models in which the stimulus information is the input while an assessment or action is the output. Models systematically differ according to their level of abstraction. Some models are specified as functional relations between high-level constructs while other models are meant to be descriptions of the actual information processing steps. Much work has been done to distinguish between different types of explanations in psychology (J. R. Anderson, 1990; Marr, 1982; Newell, 1982; Oaksford & Chater, 2007; Pylyshyn, 1984). One of the best known accounts is found in Marr (1982), who distinguishes between three levels of analysis: computational, algorithmic and physical.

Computational explanations describe what the system is doing and the logic of why that strategy is appropriate. A computational explanation specifies the ideal solution to an abstract problem. It should be mentioned that Marr's "computational" terminology has been criticized as a misnomer since at the computational level, the problem and solution are characterized mathematically and without any reference to computations (J. R. Anderson, 1990). For this reason, some prefer to describe these as "functional" explanations.

Algorithmic explanations are more specific, detailing the representations and the transformations used to execute the function. Historically, most work in cognitive science has occurred at the algorithmic level. The physical or implementational level concerns how the process is implemented by the underlying material architecture. So for human cognition, the physical explanation is in terms of neural processing.

Marr's work broadly describes different types of explanations but, as Sloman and Fernbach (2008) observe, it does not offer specific guidance for how to construct models of cognition. In particular, it does not specify which of the three levels take precedence. Should one begin with a

functional analysis of the task, or should one first focus on describing the operations actually being performed?

With his influential method of “rational analysis,” Anderson (1990, 1991a) does propose a model building procedure. He is unequivocal about where to start: one should begin with an analysis of the task that the cognitive system is trying to solve. A foremost requirement is the precise characterization of the inferential problem. In practice, this means that rational analysis nearly always starts with computational level explanations. A guiding assumption of the rational approach is that cognitive processes are well-adapted to their environment of application. Accordingly, cognitive processes should produce near optimal solutions to the particular tasks that they face. Anderson’s claim is that rational analysis is the best method for discovering and characterizing these optimal solutions.

Rational analysis has antecedents in “ideal observer theory,” an influential a research approach in the study of psychophysics. For a given sensory task, an ideal observer gives the optimal performance that can be achieved for a specific set of stimulus inputs and processing constraints (Geisler, 1989, 2011; David M. Green, 1960; David Marvin Green, 1966; Swets, Tanner, & Birdsall, 1961). Anderson’s rational analysis can be understood as an attempt to generalize ideal observer theory beyond basic sensory tasks to higher level cognitive processes, such as memory, categorization, and causal inference.

Anderson (1990, 1991a) presents the general steps of rational analysis. First, one must specify the goals of the cognitive system. Next is to give an account of the capacities available for pursuing these goals. Anderson argues that only minimal assumptions should be made concerning cognitive resources. Specifically, the assumptions should only rule out strategies that require search over a vast solution space. He justifies this position by arguing that human cognition is extremely plastic, so almost any function should be considered as a candidate explanation. Finally, one creates a formal model of the environment to which the system is adapted.

With a description of goals, capacities and the environment in place, it will then be possible to derive an optimal solution for the task so long as the solution space is "well behaved" (e.g. if the solution space is convex). The optimal solution will describe how psychological inputs (sense data) should be transformed into outputs (behavior) according to some function. Importantly, the rational model is not intended as a description of the underlying psychological process or mechanism. As such, model complexity is not viewed as problematic for the rational approach.

To provide a simple example, suppose we wish to model the behavior of a rat. The goal is to travel as quickly as possible from its starting position A to some food source B. The environment is a flat plane. Then we can derive the optimal solution: a straight line path from A to B.

So a description of an agent, its goals, and the environment allows for the derivation of a model. Yet on Anderson's (1990, 1991a) view, the task is still not complete. Anderson proposes rational analysis as an iterative procedure. After the initial solution is derived, model predictions are compared to behavior. If there are discrepancies, then the model is revised by revisiting the first steps of the model-building process. Amendments may be made to assumptions about agent goals and capacities, or the structure of the environment, in order to find a model with better empirical performance.

Rational models are sometimes described as normative, though this identification is contentious and can be somewhat confusing. A normative explanation shows the best way to perform some task, so it determines how one ought to proceed for a given problem. A rational model will be normative when its description of the environment holds and when it correctly describes the inferential goal of the agent. As will be seen at length, establishing a faithful representation of the environment and the agent's goals is a very difficult task.

From its inception, rational analysis has been paired with Bayesian inference (J. R. Anderson, 1990). This is a natural combination. Many problems in cognition require inductive inference. Bayesian inference, given certain assumptions, can be shown to be normative for such problems. As Bayesian methods have increased in popularity, the connection between rational analysis and Bayesian models has become stronger. More recent model building accounts have the Bayesian framework taking precedence (Griffiths, Kemp, & Tenenbaum, 2008; Griffiths & Tenenbaum, 2006). Accordingly, many researchers now refer to a general "probabilistic" or "Bayesian" approach to the study of cognition.

Thus, computational models are principally concerned with the "what" and "why" questions of cognition. Algorithmic or mechanistic models, in contrast, emphasize the "how" questions. They are meant to be descriptions, or close approximations, of the actual psychological processes employed. Algorithmic models are typically formulated to explain specific sets of experimental data. The upshot is that they necessarily provide good descriptions of behavior.

A guiding assumption of the algorithmic approach is that simple operations are better candidates for psychological processes, or close proxies thereof. The preference for simple models

is driven, in part, by an emphasis on the cognitive agent's computational limitations. Memory, attention, time, and processing power are all finite and costly resources. Accordingly, preferred algorithmic models are predictively strong while placing minimal demands on cognitive resources. By making resource considerations central, the algorithmic approach clearly departs from rational analysis.

An algorithmic model, like a rational model, can be thought of as a solution to an optimization problem: the optimum maximizes predictive power while minimizing psychological complexity. The trouble is that neither predictive power nor psychological complexity are well-defined. Predictive power is not defined since algorithmic models are not explicit about the inferential target. In addition, there are no clear standards for assessing psychological complexity, or equivalently, the costs and constraints imposed by the learner's cognitive infrastructure (Danks & Eberhardt, 2011). The upshot is that the "algorithmic approach" has not been explicitly codified like Anderson's rational analysis. Instead, it refers to a general research orientation that emphasizes psychological mechanism.

Rational and algorithmic approaches are each concerned with different aspects of cognition, so it is unsurprising that they each have their own merits. A strength of rational analysis is in its precision. The explicit assumptions of rational models allow for models to be clearly distinguished according to their theoretical commitments. But since rational models are almost always computational, an additional challenge is to identify plausible mechanisms that can execute rational strategies. The algorithmic approach, beginning with mechanism, faces the inverse problem of explaining why the mechanism performs as it does. Ideally, an explanation of a cognitive process would span the computational and algorithmic levels, incorporating the merits of each. To date, there has not been much work connecting rational and mechanistic models of behavior (for exceptions, see attempts by Sanborn, Griffiths & Navarro (2010) or Oaksford & Chater (2010)).

A major goal of this dissertation is to work towards a tighter integration of rational and algorithmic models of causal learning. This is the primary focus of chapters 1 through 4. Progress may be achieved with current experimental paradigms and existing models of causal learning, but certain barriers persist. So in Chapter 5 an attempt to expand the study of causal learning is made with the formulation of a latent variable framework. The outline of the dissertation is as follows:

Chapter 1 reviews the causal learning experimental paradigm and discuss a number of influential models that have been proposed in the area.

Chapter 2 closely examines the mechanics of Bayesian models of causal learning. I then review the larger debate on the use of Bayesian models to describe cognitive behavior. I argue that the causal learning task is amenable to a Bayesian analysis, but that the evidence is equivocal on whether people's behavior is Bayesian.

Chapter 3 investigates the weighted ΔP rule as an alternative to Bayesian models of causal learning. I propose a novel strategy for bridging levels of analysis, which I use to connect the algorithmic weighted ΔP rule to the computational causal power model. This allows for an investigation of weighted ΔP as an estimator of causal strength. I show that weighted ΔP is a low variance estimator of causal strength in the context of parameter and model uncertainty. I also demonstrate that an iterative version of weighted ΔP converges to causal power.

Chapter 4 closely examines causal learning data in order to better distinguish between the weighted ΔP model and Bayesian models of causal learning. I document the existence of a "deterministic bias", which is generally incompatible with Bayesian models of strength estimation. In contrast, the weighted ΔP model does predict the deterministic bias for the two conditions in which it is most commonly observed.

Chapter 5 takes a general view of rational models of causal learning. I construct a latent-variable framework for causal models, which is called *capacity and response probability* (CARP). On this approach, a response function maps latent causal capacity to an effect probability. Different functions correspond to different assumptions about the causal system. The two primary rational models, the ΔP rule and causal power, are shown to be two particular models along a continuum of models. I then speculate on how CARP may be used to measure actual causal environments. I also show that the latent-variable formalism allows for new intuitions and insights about causal learning.

Chapter 6 concludes the dissertation.

Chapter 1.

Causal learning models and evidence

1.1 Causal learning paradigm

The problem of elemental causal induction concerns the relationship between a single binary cause and a single binary effect (Griffiths & Tenenbaum, 2005). The focus of this dissertation is on models that explain judgments for this basic relationship. Clearly, many factors influence causal judgments, such as temporal and spatial contiguity, and domain-specific knowledge (Einhorn & Hogarth, 1986). Research on elemental causal induction attempts to hold these other factors constant in order to isolate the influence of contingency.

For elemental causal induction problems, learning information can be characterized with a 2x2 contingency table (Table 1.1). The four frequencies are often represented by the variables *a*, *b*, *c* and *d* beginning in the top-left cell and moving down by rows. In causal learning experiments the contingency table outcomes are presented in one of three formats. The *summary format* simply presents the table as the experimental stimulus, or as a graph that conveys the same information. For example, Ward and Jenkins (1965) presented information in a contingency table while Buehner and Cheng (1997) used a pair of pie-charts to represent the c^+ and c^- conditions.

The *sequential* or *online format* presents outcomes trial-by-trial. Arkes & Harkness (1983) assert that trial-by-trial presentation provides a closer analog to how people estimate contingency in natural settings in which they must rely on their memory.

Table 1.1. A 2x2 contingency table

	Effect Present (e^+)	Effect Absent (e^-)	
Cause Present (c^+)	$N(e^+, c^+)$	$N(e^-, c^+)$	$N(., c^+)$
Cause Absent (c^-)	$N(e^+, c^-)$	$N(e^-, c^-)$	$N(., c^-)$
	$N(e^+, .)$	$N(e^-, .)$	$N(., .)$

The *list format*, found in more recent experiments, shows the individual cases simultaneously in a single array or list. This format reduces memory demands while preserving the structure of individual learning trials. Lu, Yuille, Liljiholm, Cheng & Holyoak (2008) favor simultaneous presentation as they believe that memory effects should be minimized in order to isolate causal inference. It is questionable, though, whether the same reasoning strategy is used in both the sequential and simultaneous formats. For instance, experiments by Perales and Shanks (2008) suggest that the list format does not just reduce cognitive demands, but also changes the reasoning strategy as well.

After presentation of the learning data, participants are asked to make a causal strength judgment. Traditional wording of the question is along the lines of "judge the extent to which the cause *C* produces the effect *E*" (Perales & Shanks, 2008). The phrasing of the question has been a point of controversy, which will be discussed further below.

1.2 Rational models

Several rational models of causal learning have been proposed, each giving different “optimal” solutions for the same learning environments (see the introductory chapter for a description of rational analysis). Only one can truly be optimal, so it is instructive to examine the commitments of a particular model. The argument for a given rational model rests on two premises: 1) The given model will be the best at discovering causes and their magnitudes as they exist in the real world and 2) human causal judgments are well-adapted, so they should conform to rational inference. All rational models share the second premise, so it is on the first where the disagreement resides. The first premise has been argued on *a priori* grounds, as to be seen momentarily. Additionally, the two premises taken together imply specific predictions about human judgment. Section 1.6 reviews work that has attempted to empirically distinguish between various models of causal learning.

The discussion of rational models uses notation that will be found throughout the dissertation. This is summarized in Table 1.2 below.

Table 1.2. Notation for causal learning models

Notation	Definition
E	effect (binary)
C	candidate cause (binary)
B	background cause or context (binary)
e^+	effect occurs
e^-	effect does not occur
c^+	candidate cause is present
c^-	candidate cause is absent
b^+	background cause is present
b^-	background cause is absent
w_0	causal strength of *B* (edge weight from *B* to *E*)
w_1	causal strength of *C* (edge weight from *C* to *E*)
w_T	combined or "total" causal strength of *C* and *B*

1.2.1 The ΔP rule

The ΔP rule is a prominent rational model of causal learning. The counts from Table 1.1 can be used to find the conditional probabilities $P(e^+|c^+)$ and $P(e^+|c^-)$. The ΔP rule is just the difference in the conditional probabilities:

$$\Delta P = P(e^+|c^+) - P(e^+|c^-) \tag{1.1}$$

Many have argued that ΔP is a normative measure of causal strength (Allan, 1980; Jenkins & Ward, 1965; Ward & Jenkins, 1965). The ΔP rule is also attractive in its simplicity. The model requires only the simple operations of frequency encoding (Hasher & Zacks, 1984) and forming a difference. The ΔP model does not require that a causal direction be specified (Shanks, 1995).

Another appeal is its relationship to the well-known Rescorla-Wagner (1972) model, which is elaborated in Section 1.5.1 below.

Patricia Cheng (1997) argues against ΔP as a normative model, asserting that it only measures association, not causation. To give one example, white hair is reliably associated with heart disease, but this does not imply that white hair causes heart disease. Association may instead be due to a shared common cause. In this example, white hair and heart disease share the common cause of aging.

A related criticism of ΔP concerns how the strength estimates depend on the context. Suppose, for example, that country A has a very high infant mortality rate of 20%. In addition, suppose a rare genetic disorder has been discovered in country A. Infants who test positive at birth have a mortality rate of 100%. For country A, the ΔP rule returns a strength estimate of 80% for the disease as a cause of death. Now imagine country B with an infant mortality rate of 1%. The ΔP model predicts that 81% of infants who test positive in country B will die from the disease. Intuition suggests this prediction is much too low. Most people would probably predict a country B mortality rate of around 100%.

1.2.2 Power PC theory

The examples above suggest that people distinguish between covariation and causation. Cheng constructs a formal distinction between association and causation with her seminal power PC theory, which is meant to replace ΔP as the normative model of causal inference. On Cheng's account, people interpret covariation information with respect to a framework of beliefs. These beliefs include a notion of "causal powers" that determine how causes influence their effects. The goal of causal induction is to estimate these powers.

More specifically, suppose that a person is interested in the relationship between a candidate cause *C* and its purported effect *E*. One assumption of the power PC model is that if the effect *E* occurs then it must have been caused by either the observed candidate cause *C* or by some background cause *B* (Figure 1.1, Graph 1). The background or context *B* comprises all other causes that might possible influence the effect. The edge weights w_0 and w_1 in Graph 1 represent the causal strengths or "powers" of the background cause *B* and the candidate cause *C*, respectively.

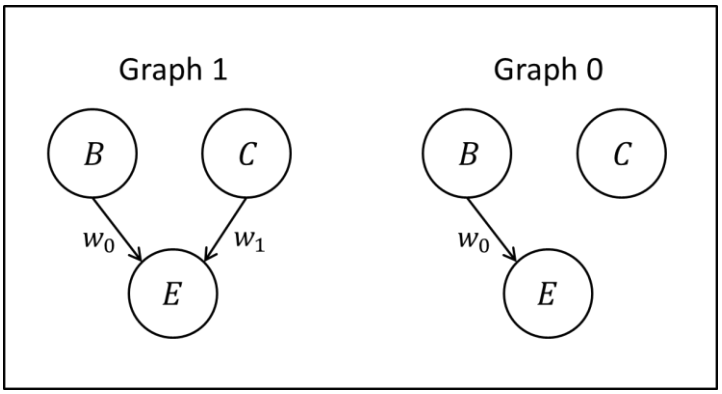


Figure 1.1. Directed graphs with B representing background variables, C as the candidate cause and E as the effect of interest.

Holyoak and Cheng (2011) list the four primary assumptions of causal power as:

1) *B* and *C* influence the effect *E* independently.
2) *B* could produce *E*, but not prevent it.
3) Causal powers are independent of the frequency of occurrences of the causes.
4) *E* does not occur unless it is caused.

The first two assumptions are taken as default hypotheses adopted by the judge, which may be revised in the face of conflicting evidence. Assumption 1 is sometimes referred to as the "no confounding" assumption. Assumptions 3 and 4 are argued to be fundamental to causal inference. From these assumptions, Cheng derives her power PC model of causal strength. Appendix A argues that Cheng's assumption 3 is problematic. In order to resolve this difficulty, the appendix presents a modified derivation that only requires assumptions 1, 2, and 4.

An additional detail is that the background causes that comprise *B* are not observed (or they ignored). Indeed, if they were observed then these factors would constitute additional candidate causes. A practical assumption is often made to facilitate accounting for the influence of the context. That is, the background cause *B* is assumed to be always present so that $P(b^+) = 1$. This assumption will be adopted throughout this dissertation unless explicitly stated otherwise. As a consequence, some notation will often be abbreviated. Namely, the truncated expressions $P(e^+|c^+)$ and $P(e^+|c^-)$ should be interpreted, respectively, as $P(e^+|c^+, b^+)$ and $P(e^+|c^-, b^+)$ unless specifically stated otherwise.

The power PC model requires first determining the candidate cause's direction. The sign of ΔP, which can be inferred from the observed data, determines causal direction (Buehner, Cheng,

& Clifford, 2003). When ΔP is positive the cause is considered "generative" of the effect, and generative causal power (henceforth, abbreviated as "generative power") is given by:

$$\text{generative causal power} = \frac{\Delta P}{1 - P(e^+|c^-)} \tag{1.2}$$

From (1.2) we can see that for a fixed ΔP, generative power will return increasing judgments of causal strength as the base rate of the effect $P(e^+|c^-)$ increases. When the effect always occurs $P(e^+|c^+) = P(e^+|c^-) = 1$ and generative power is undefined with a denominator equal to 0. The intuition is that if the effect always occurs there is no opportunity for the cause to demonstrate its influence and so no inference can be made.

An interesting bit of trivia is that generative power is given by the same equation as Cohen's kappa (Cohen, 1960). Kappa is used to measure agreement between two raters. In the equation for kappa, $P(e^+|c^+)$ corresponds to the observed agreement among raters and $P(e^+|c^-)$ corresponds to the hypothetical probability of chance agreement.

With a negative ΔP the candidate cause is "preventive", and the power PC model takes the form:

$$\text{preventive causal power} = \frac{-\Delta P}{P(e^+|c^-)} \tag{1.3}$$

Similar to the generative case, preventive power is undefined when $P(e^+|c^-) = 0$.

Causal power is a context-independent measure of causal strength, wherein context refers to the varying assemblage of background causes (Cheng, 2000). This is achieved in the denominator of (1.2) and (1.3), which normalizes ΔP by the base-rate of the effect $P(e^+|c^-)$. A causal power gives the probability that a given cause "working in isolation" will produce the effect. Cheng's model is inspired, in part, by Nancy Cartwright's (1989) work on causality. The connection to Cartwright's work will be explored in more depth in Chapter 5. Cartwright (2007) gives extensive treatment of the conception of causal powers and what it means for a cause to be isolated.

1.3 Causal graphical models

Rational models of causal learning have benefitted from connections made to work in computer science and statistics. Clark Glymour (1998, 2000) showed that the causal power model could be

represented using the formalism of graphical causal models. These models consist of a set of nodes or variables and a set of directed edges between nodes. Directed edges travel from "parent" nodes to their "child" nodes. In graph 1 of Figure 1.1, *B* and *C* are the parents of *E*, and so *E* is the child of *B* and *C*. A graph is "causal" by virtue of a specific set of assumptions that have been formulated so that the directed links are meant to represent causal relationships (Glymour, 2002; Pearl, 2009; Spirtes, Glymour, Clark, & Scheines, 1993).

The structure of a graph is given by the edges between the variables. Graphical structure implies general information about the joint probability distribution, namely the pattern of dependence between variables. However, a specific joint distribution is determined by a graph's parameterization or functional form. The parameterization specifies how the variables influence one another. Different assumptions about causal relationships can be expressed using various functional forms.

The w_0 and w_1 in Figure 1.1 are edge weights, which are used to describe the parameterization of a graph. Different parameterizations embody different causal hypotheses, and the edge weights emerge as measures of causal strength in the context of a specific model. In Graph 1 there are two edges into the effect node, and so the parameterization must account for how *C* and *B* interact to produce the effect *E*. In particular, Graph 1 has $P(e|c,b) = f(w_0, w_1)$, where $f(.)$ is some unspecified function mapping into the $[0,1]$ interval. Below it will be shown how different assumptions concerning *B* and *C* lead to different choices of $f(.)$. In contrast, Graph 0 only has one edge from *B* into *E*, and the parameterization only requires specifying the conditional probability $w_0 = P(e^+|c^-, b^+)$.

Graph 1 in Figure 1.1 is known as a common-effect causal graph. From the properties of causal graphs one can conclude from the graph that *C* and *B* are statistically independent. This is because no edges directly connect *C* and *B* and because they have no common parents. Tenenbaum and Griffiths (2001) show that the ΔP and causal power models correspond to different functional forms on a common-effect graph. In particular, these models give maximum-likelihood estimates (MLEs) for the value of the edge weight w_1 under different parameterizations of Graph 1.

1.3.1 Linear parameterization

The ΔP rule gives the MLE for causal strength w_1 assuming a linear parameterization of Graph 1 (Griffiths & Tenenbaum, 2005). The linear parameterization assumes the probability of the effect in the presence of the background *B* alone is w_0 and that the cause *C* changes this probability by a constant amount:

$$P(e^+|c,b) = w_0 \times b + w_1 \times c \tag{1.4}$$

Where c^+ means $c = 1$, c^- means $c = 0$, and e^+, e^-, b^+ and b^- are defined correspondingly. Note we must also have $w_0 + w_1 \in [0,1]$ to obtain a legal probability. It is straightforward to see that ΔP equals the edge weight w_1 under the linear parameterization. Since $w_0 = P(e^+|c^-, b^+)$, we obtain:

$$P(e^+|c^+, b^+) = P(e^+|c^-, b^+) + w_1$$
$$w_1 = P(e^+|c^+, b^+) - P(e^+|c^-, b^+) = \Delta P$$

When ΔP is computed from a sample, as from Table 1.1, the sample proportions are used to estimate the conditional probabilities with $\hat{P}(e^+|c^-, b^+) = \frac{N(e^+,c^-)}{N(e^+,c^-)+N(e^-,c^-)}$ and $\hat{P}(e^+|c^+, b^+) = \frac{N(e^+,c^+)}{N(e^+,c^+)+N(e^-,c^+)}$. Since the sample proportions are the MLEs for the corresponding population probabilities, ΔP is the MLE for the causal strength parameter w_1.

1.3.2 Noisy-OR parameterization

The Noisy-OR parameterization involves a different set of assumptions for how *C* and *B* interact. First, both causes are assumed to be generative, meaning that they can only increase the probability that the effect occurs. The other key assumption is that when both *B* and *C* are present, they each have an independent opportunity to produce the effect. This set of assumptions yields the Noisy-OR model:

$$P(e^+|c,b) = 1 - (1 - w_0)^b (1 - w_1)^c \tag{1.5}$$

Again, with only the background present, $P(e^+|c^-, b^+) = w_0$. But now both causes present gives the conditional probability $P(e^+|c^+, b^+) = w_1 + w_0 - w_1 \times w_0$. Following the same strategy as above and solving for w_1 results in the expression for generative power from (1.2). If the contingency data are used to find the sample proportions $\hat{P}(e^+|c^-, b^+)$ and $\hat{P}(e^+|c^+, b^+)$, then the generative power equation with these quantities gives the MLE for the strength parameter w_1 (Griffiths & Tenenbaum, 2005).

1.3.3 Noisy-AND-NOT parameterization

If cause *C* is preventive then w_1 is the probability that *C* prevents *E*. In this case, the effect will occur if it is generated by *B* and not prevented by *C*. Independence of *B* and *C* gives the Noisy-AND-NOT parameterization:

$$P(e^+|c, b) = w_0^b(1 - w_1)^c \tag{1.6}$$

With both causes present $P(e^+|c^+, b^+) = w_0 \times (1 - w_1)$. Again, solving for w_1 gives the expression for preventive power. And using the sample proportions for the conditional probabilities in (1.6) will give the MLE for the strength parameter w_1.

1.3.4 Additional parameterizations

This causal graph formalism can also be extended to additional models. For instance, examine the parameterization:

$$P(e^+|c, b) = w_0 \times b + w_1 \times c - w_0(c \times b) \tag{1.7}$$

This model yields a causal strength of $w_1 = P(e^+|c^+, b^+) = \frac{N(e^+, c^+)}{N(e^+, c^+) + N(e^-, c^+)}$. This has been referred to by Ward and Jenkins (1965) as "Per Cent Success" and by Klayman and Ha (1987) as "+Htesting". It is an algorithmic model of inference wherein participants completely ignore the base rate of the effect. The model in (1.7) is a special case of the 1-parameter weighted ΔP model:

$$P(e^+|c, b) = w_0 \times b + w_1 \times c - k \times w_0(c \times b) \tag{1.8}$$

where $k \in [0,1]$. Though weighted ΔP can be expressed as a parameterization of a common effect graph, it can easily be shown that (1.8) makes a very poor candidate as a rational model.

1.4 Bayesian models of causal learning

Bayesian methods have been influential throughout the development of rational models, with their impact continuing to grow, especially over the last ten years. Early applications are found in the work of Anderson (1990) and Fales and Wasserman (1992). Griffiths and Tenenbaum (2005) were novel in their union of Bayesian methods to the ΔP and causal power models. Recent Bayesian models posit that human inference proceeds as if people use the learning data to form a posterior distribution for causal strength. Causal judgments are then constructed as some function of the posterior distribution. Typically, the judgment is given by the posterior expectation, though some models use only a few or even one sample from the posterior (e.g. (Vul, Goodman, Griffiths, & Tenenbaum, 2014)).

The posterior distribution is found by combining a prior distribution and a sampling model, which is also referred to as a likelihood. These two components will be briefly described in turn (see Hoff (2009) for a detailed treatment). The prior encodes beliefs about parameter values before any data have been observed. The height of the prior is used to find absolute levels of belief for different regions of the parameter space. Suppose we have a prior for the parameter θ given by a continuous probability density function $p(\theta)$. For a continuous density, the probability of θ taking any specific value $\tilde{\theta}$ is zero. However, we can find the probability that θ takes a value within some small neighborhood of $\tilde{\theta}$. If the length of the neighborhood is ϵ, then the probability is approximately $\Pr(\theta = \tilde{\theta}) \approx \epsilon \times p(\tilde{\theta})$, or the length multiplied by the height. The shape of the prior density also determines relative levels of belief. For example, $\frac{p(\theta_a)}{p(\theta_b)} = 2$ indicates that the value θ_a is considered twice as probable as the value θ_b.

The sampling model $p(D|\theta)$ describes the belief that the data D would be observed for a fixed parameter value θ. We can also think of the data as fixed and examine the sampling model as a function of θ. This relationship is described by the likelihood function, typically written as $L(\theta|D) = p(D|\theta)$. The shape of the likelihood function $L(\theta|D)$ conveys how the data inform the parameter estimates. The maximum likelihood estimate is the just the value $\hat{\theta}_{\text{MLE}}$ that maximizes the function $L(\theta|D)$ for the observed data. While the location of the peak of $L(\theta|D)$ determines

the MLE, the shape of $L(\theta|D)$ at the MLE reveals how informative the data are. Specifically, the likelihood function can be used to estimate uncertainty regarding $\hat{\theta}_{\mathrm{MLE}}$ (Cramer, 1986). If the likelihood is highly curved at $\hat{\theta}_{\mathrm{MLE}}$, then the function is changing rapidly and nearby likelihood values quickly become much lower than $L\left(\hat{\theta}_{\mathrm{MLE}}\middle|D\right)$. On the other hand, if the likelihood is nearly flat, then nearby θ values give likelihoods that are almost as good as $L\left(\hat{\theta}_{\mathrm{MLE}}\middle|D\right)$. Consequently, high curvature near $\hat{\theta}_{\mathrm{MLE}}$ is associated with low variance while a flatter likelihood corresponds to a higher variance for $\hat{\theta}_{\mathrm{MLE}}$.

With the prior and the likelihood, the posterior distribution $p(\theta|D)$ is found using Bayes rule:

$$p(\theta|D) = \frac{p(D|\theta)\times p(\theta)}{p(D)}$$

Since the probability of the data $p(D)$ does not depend on the parameter θ we can write $p(\theta|D) \propto p(D|\theta)\times p(\theta)$. This says that the posterior is proportional to the likelihood multiplied by the prior. The posterior, then, is a sort of compromise between the likelihood and the prior. In practice, it can often be difficult or impossible to find an exact expression for the posterior. For these cases Monte Carlo methods can be used to numerically approximate the posterior distribution.

Specific to the problem of elemental causal induction, the prior distribution $p(w_0, w_1)$ is for w_0 and w_1, the background and candidate causal strengths respectively. The data *D* are the frequencies from a 2×2 contingency table. The likelihood $p(D|w_0, w_1)$ can be specified as a binomial distribution. Thus, one must assume a parameterization for how *B* and *C* combine in their strengths w_0 and w_1 to give the binomial probability for each of the four trial types. The parameterization is referred to as the *generating function* in the Bayesian context. Griffiths and Tenenbaum (2005) show that generating functions can be chosen to reflect either ΔP or causal power as the underlying model. Specifically, they show that the linear generating function corresponds to the ΔP rule. And for causal power, the Noisy-OR and the Noisy-AND-NOT are the generating functions for generative and preventive causes, respectively.

In their Bayesian model of "causal support", Griffiths and Tenenbaum (2005) emphasize the distinction between causal strength and causal structure. The causal support model formulates judgment as a Bayesian decision about whether a causal relationship exists. That is, causal support

gives Bayesian confidence for the hypothesis that strength w_1 is not equal to 0. Equivalently, this is the hypothesis test on whether a set of observations were generated from Graph 1, in which *C* causes *E*, or were generated from Graph 0 in which *C* does not influence *E*. One can find the posterior probability of each hypothesis given the data and compare them by taking their ratio. If prior belief for the two hypotheses is equal, then the posterior ratio is just equal to the likelihood ratio of $\frac{P(D|\text{Graph 1})}{\text{P(D|Graph 0)}}$. Finding the likelihoods takes some work, especially for $P(D|\text{Graph 1})$, as it requires integration over the (w_0, w_1) parameter space. In their appendix Griffiths and Tenenbaum (2005) provide an algorithm to approximate this integral.

More recent models assume that people do in fact assess causal strength, but that they use Bayesian inference for their judgments. On this perspective, causal learning is akin to parameter estimation for the causal graph edge weight w_1. These models assume that people implicitly draw a random sample of values from their posterior distribution of w_1 in order to form strength judgments (Lu, Yuille, Liljiholm, Cheng, & Holyoak, 2007). For this reason, the posterior expectation of w_1 is used to make model predictions.

One Bayesian model of causal strength assumes a joint uniform prior over the background and focal strengths so that $p(w_0, w_1) = 1$ for all $w_0, w_1 \in [0,1]\times[0,1]$. The uniform prior is popular since it is agnostic in belief: any pair of values (w_{0i}, w_{1j}) is viewed as equally probable as any other pair (w_{0h}, w_{1k}).

Lu et al. (2006, 2007, 2008) claim that people have generic priors for sparse and strong (SS) causes. The SS prior distribution reflects a preference for sole strong causes. For the generative case this entails a prior distribution with peaks over (w_0, w_1) at (0,1) and at (1,0), which respectively attributes *C* or *B* as the sole strong cause. Thus, the candidate and background causes compete as explanations for the effect. The model contains a single free parameter α, which determines how strongly SS priors are preferred. An $\alpha = 0$ gives a uniform distribution and no preference for an SS prior. As α becomes more positive, the heights of the peaks over (0,1) and (1,0) grow, reflecting a stronger belief in sparse strong causes. Lu et al. (2007) found that $\alpha = 5$ provides a good fit to human causal judgments.

Yeung and Griffiths (2015) investigate Bayesian models also using causal power generating functions. However, instead of assuming a particular prior distribution they seek to estimate its shape from data. They use a Markov-chain Monte Carlo (MCMC) technique of iterated learning

(Griffiths, Christian, & Kalish, 2008; Griffiths & Kalish, 2007). If model assumptions hold, then once the process has converged they can obtain many samples from peoples' prior distributions. In the paper they use a smooth non-parametric surface to fit the prior samples. The surfaces they obtain look markedly different from the shape of SS priors. It is difficult to confidently draw conclusions from the fitted surfaces since they are non-parametric. However, the general shape appears to reflect a prior belief in strong causes, with the density much higher for large values of w_1 (both for generative and preventive causes). In contrast to the SS priors, the w_1 and w_0 look to be relatively independent. That is, the shape of the w_1 distribution is fairly similar at different levels of the w_0 distribution.

To achieve predictions, all of the above Bayesian models require integration over the posterior for which there is no analytic solution. Instead, the integrals must be approximated numerically. It is important to emphasize that the Bayesian models, like all rational models, make no claims about representation or implementation. As such, technical challenges of finding model predictions, such as the representation of a complex posterior distribution, are not viewed as problematic for the Bayesian approach. The chief concern is whether the Bayesian predictions capture key patterns in judgment. That said, there has been increasing emphasis on mechanistic models that can approximate Bayesian inference. The details of this research will be discussed in the next two chapters.

1.5 Algorithmic models

An abundance of algorithmic models of causal judgment have also been proposed, and this continues to be an active area of research. Perales and Shanks (2007) distinguish algorithmic models as either *rule-based* or *associative*. Rule-based models assume that the learner explicitly stores all outcome frequencies, which are then combined according to some rule. Associative models claim that the causal strength judgment is formed via the incremental accumulation of association strength. Rule-based and associative models are now examined in turn.

1.5.1 Rule-based models

A number of proposed models are just simple rules applied to the frequencies of Table 1.1. These include "Cell A" strategies wherein causal judgment is simply an increasing function of the cell *a*,

or the $N(e^+, c^+)$ frequency (Nisbett & Ross, 1980; Smedslund, 1963). The "A versus B" rule uses the entire top row of Table 1.1 and predicts strength judgments from the difference $N(e^+, c^+) - N(e^-, c^+)$ (Arkes & Harkness, 1983; Kao & Wasserman, 1993). The Per Cent Success rule from above also restricts attention to the top row only.

The ΔD or sum of diagonals rule is another model that has received extensive study (Arkes & Harkness, 1983; Inhelder & Piaget, 1958; Jenkins & Ward, 1965; Kao & Wasserman, 1993; Shaklee & Tucker, 1980; Shanks, 1987; Wasserman, Chatlosh, & Neunaber, 1983; Wasserman, Dorner, & Kao, 1990). The ΔD rule uses all entries from the contingency table by predicting judgments with:

$$\begin{aligned} \Delta D &= [N(e^+, c^+) + N(e^-, c^-)] - [N(e^-, c^+) + N(e^+, c^-)] \\ &= [\mathrm{a} + \mathrm{d}] - [c + b] \end{aligned}$$

White's (2003) pCI rule is closely related to the sum of diagonals rule. White normalizes ΔD so that it ranges from -1 to 1. He does this be specifying $\mathrm{pCI} = \frac{\Delta D}{N(.,.)}$, so that it is normalized by the sample size. When the number of observations in the c^+ and c^- trials are equal, pCI is identical to the ΔP rule.

Finally, the ΔP rule, and even causal power, can be considered rule-based algorithmic models as they each provide a rule for combining cell frequencies to arrive at a causal strength judgment. Whether these models are rational, algorithmic, or both or neither, ultimately depends on theoretical commitments of the researcher. For instance, a researcher may claim that typical causal systems (for binary causes and effects) are best characterized by a linear parameterization over a common-effect causal graph. The very same researcher may also claim that when people are presented with learning trials they use it to construct internal representations $P(e^+|c^+)$ and $P(e^+|c^-)$, and then they take the difference of these conditional probabilities. For this researcher, then, the ΔP rule serves as both a rational and an algorithmic model. Whether the researcher is justified in this belief is a separate matter. Indeed, there are strong reasons to doubt both claims.

The rule-based models discussed thus far are *unweighted* since they use the objective frequencies from the cells of the contingency table. Yet people do not appear to give equal weight to the different trial types (Allan, 1993; Einhorn & Hogarth, 1986; Kao & Wasserman, 1993; Levin, Wasserman, & Kao, 1993; Ward & Jenkins, 1965; Wasserman et al., 1990). In order to

improve descriptive power, numerous *weighted* models have been proposed. These models have one or more free parameters that are estimated from human data.

The free parameters, or weights, are believed to reflect a number of psychological factors. For instance, they could represent differences in attention that are driven by the beliefs of the judge or the salience of the stimuli (Busemeyer, 1991). Further, weights are often interpreted as distortions of the normative model. As a result, no computational analysis of the weights is attempted beyond a simple comparison with the normative standard.

The weighted ΔP rule forms a class of causal judgment models. Several versions of weighted ΔP have been posited by various researchers (J. R. Anderson & Sheu, 1995; Busemeyer, 1991; Kao & Wasserman, 1993; Lober & Shanks, 2000; Wasserman, Elek, Chatlosh, & Baker, 1993) The two parameter weighted ΔP rule is:

$$\mathrm{w}\Delta P = k_1 P(e^+|c^+) - k_2 P(e^+|c^-)$$

where k_1 and k_2 are the weights for the two conditional probabilities. A typical finding is that $k_1 > k_2$ gives a good description of observed data.

Another well-known model is the Evidence Integration or EI rule, which is a weighted version of the pCI rule. The EI rule has a total of four weights, one for each cell in Table 1.1 (Perales & Shanks, 2007).

1.5.2 Associative models

Many researchers consider rule-based models to be implausible as proper algorithmic models. They believe rule-based models assume processes and representations that are too complex for an algorithmic implementation. Shanks (1995), for instance, is skeptical that responses are derived from an explicit version of the ΔP rule, even though he does favor ΔP as descriptive of judgments. For subjects to use an explicit version of ΔP, they would need to continuously maintain probability estimates of $P(e^+|c^+)$ and $P(e^+|c^-)$, which would be updated and contrasted on a trial-by-trial basis. Causal power is even more complex since it includes the additional step of norming the contrast by the base rate of the effect.

Associative models are intended to be more faithful representations of psychological processes. As such, they are specified to make minimal memory and computational demands on

the learning agent. Skepticism about an animal's ability to sort and count events over large blocks of time motivated, in part, the Rescorla-Wagner model (R-W model) of associative learning (Rescorla & Wagner, 1972). To wit, Wagner (1969) states that sensitivity to correlation between cue and reinforcement is "…not as a consequence of some complex experiential contrast between the probabilities…but rather as a consequence of the resulting trial-by-trail signal value…" (p.33). While originally developed to explain animal learning, the R-W model has also been proposed as an account of human contingency learning (Shanks, 1995; Shanks & Dickinson, 1987; Wasserman et al., 1993). On this account, causal reasoning is just a manifestation of associative learning.

The R-W model only requires the representation of association strengths which are incremented up or down on a trial-by-trial basis. This eliminates the need to maintain a memory of all previous trials. In the R-W model, causal strength is equated with association strength. The model assumes an always present context or background cause C_0 (denoted above as B). A cause C_i that appears on trial t will experience a change in its association strength V_i. This change is given by:

$$\Delta V_i^t = \begin{cases} \alpha_i \beta^+ \left[\lambda - \sum V_j\right] & \text{if both } C_i \text{ and } E \text{ appear in trial } t \\ \alpha_i \beta^- \left[0 - \sum V_j\right] & \text{if } C_i \text{ appears and } E \text{ does not appear in trial } t \end{cases}$$

The parameter α_i describes the salience of the cause C_i. The β^+ and β^- are the respective learning rates for trials when the effect E does or does not occur. The λ gives the maximum possible association strength. For human contingency learning λ is typically set to 1.

The $\sum V_j$ is the sum of association strengths for all causes C_j that also appear in trial t. In the elemental causal induction task, this will only include two quantities: the association strength V_1 for the candidate cause C_1 and the strength V_0 for background cause C_0. The sum of association strengths $\sum V_j$ can be thought of as the expectation for trial t. The difference $\lambda - \sum V_j$ or $0 - \sum V_j$ gives the deviation from expectation. The difference is then used to update the association strengths in a process called "error correction." Thus, the R-W algorithm learns incrementally in response to prediction error. At equilibrium the expected change in association strength is 0, or $E[\Delta V_i^t] = 0$.

Different assumptions about the learning rates give different versions of the R-W model. The restricted version sets $\beta^+ = \beta^- = \beta$, while the unrestricted version allows the learning rates to differ. With a single candidate cause the restricted R-W will reach an equilibrium equal to the ΔP rule (Chapman & Robbins, 1990); see Cheng (1997) and Danks (2003) for more general results. Hence, the R-W model gives one possible algorithmic implementation of the ΔP rule. If one assumes judgments have reached equilibrium, the two models give identical predictions in the aggregate.

It has been argued that associative models are applicable only to trial-by-trial presentation formats (Kao & Wasserman, 1993; Lober & Shanks, 2000). Association weights are built gradually using single increments and decrements. Such a process is impossible with the summary format since individual trials are not available. With the list format, sequential processing is possible, leading some to argue that associative models are still appropriate (e.g. Buehner et al., 2003). This argument seems implausible, though, when one considers how much more quickly participants proceed through list format conditions relative to trial-by-trial designs.

While the Rescorla-Wagner model is the most influential associative account, others have also been proposed. Busemeyer (1991) describes how to incrementally update weighted-averaging models. Pearce (1987) proposes an associative model of stimulus generalization. Researchers have also attempted to adapt Pearce's model to human causal learning (Baker, Vallée-Tourangeau, & Murphy, 2000; Perales & Shanks, 2003; Vallée-Tourangeau, Murphy, Drew, & Baker, 1998).

1.6 Empirical findings

There has been much empirical work investigating models of causal learning. A number of experiments have contrasted the predictions of causal power with the ΔP model. A standard strategy has been to test conditions in which causal power predictions are held constant while ΔP is varied, or vice-versa. Evidence in support of the two models has been mixed, and there have been various proposals to resolve findings that disagree with one's preferred model.

1.6.1 Early investigations of causal power

Buehner and Cheng (1997) compared causal power and ΔP in a number of experiments. In their generative experiment they tested conditions in which ΔP was equal to a positive constant while

the base rate of the effect $P(e^+|c^-)$ varied. Consistent with causal power, they found that human causal judgments, as measured by the standard probe, increased with increasing $P(e^+|c^-)$. Similarly, in their preventive experiment they found that when ΔP was equal to a negative constant, human judgments increased with decreasing $P(e^+|c^-)$.

In six experiments Lober and Shanks (2000) further investigated causal power. They used an online format for the first three experiments and a summary format for the latter three. Lober and Shanks held causal power constant and varied ΔP in four of the experiments. In all of these they found that judgments significantly increased with increasing ΔP. Judgments conformed to causal power predictions only in experiment 3, which varied power while keeping ΔP constant. Lober and Shanks ventured that an unrestricted R-W model could also explain these findings, which was applicable since experiment 3 used an online format.

To explain the influence of the base rate, Lober and Shanks (2000) specify an unrestricted R-W model that allows unequal learning rates across the e^+ and e^- trials with $\beta^+ < \beta^-$. For a fixed ΔP, this unrestricted R-W model predicts larger absolute magnitudes of judged strengths as the base rate of the effect $P(e^+|c^-)$ increases. Note that this is the same ordinal prediction as causal power. Lober and Shanks argue that the for experiments using a generative context, the $\beta^+ < \beta^-$ ordering is reasonable since participants will expect the effect to occur. The absence of the effect, then, will cause greater surprise and attention, which is reflected in the learning rates. A parallel argument can be made for experiments using a preventive context for why the occurrence of the effect will be more surprising, implying the reverse ordering of $\beta^- < \beta^+$. This implies larger judged magnitudes with a decreasing base rate, again the same as causal power. Taken together, Lober and Shanks conclude that associative learning can explain key findings from studies of causal judgment.

Buehner et al. (2003) explored whether unequal learning rates in the R-W model could, in fact, account for causal judgments. They performed experiments with a neutral background context that included both positive and negative contingencies. Since a single context was used, they argue that a single set of R-W learning rates should apply across the positive and negative ΔP conditions. The unrestricted R-W model should then predict the same influence of the base rate across positive and negative contingencies. Buehner et al. (2003) found that the effect of the base rate still reversed across positive and negative ΔP, in agreement with causal power. For this reason, among others, Buehner et al. (2003) argue that causal power is the best descriptive model and the unrestricted

R-W model should be rejected. Indeed, Perales and Shanks (2003) experiment 1 results were also inconsistent with the unrestricted R-W model.

While some results have supported causal power and others have favored ΔP, there are certain findings that contradict both models. One serious problem is that human judgments are significantly non-zero for noncontingent conditions (i.e. conditions with $\Delta P = 0$). In fact, there is a reliable pattern to these judgments: when participants are primed with a generative cause and $\Delta P = 0$, judgments of strength increase with an increasing base-rate of the effect. A similar but opposite pattern consistently emerges when participants are primed with a preventive cause—now strength ratings increase with a decreasing base rate of the effect. The phenomenon of nonzero judgments in $\Delta P = 0$ conditions has been referred to as the "frequency illusion" or the "outcome-density bias". The frequency illusion has been consistently observed in many studies (e.g. Allan & Jenkins, 1983; Buehner & Cheng, 1997; Jenkins & Ward, 1965; Shanks, Lopez, Darby, & Dickinson, 1996).

The upper left panel of Figure 1.2 shows data from Buehner and Cheng's (1997) generative component of experiment 1 (data obtained from Buehner et al. (2003)). The frequency illusion is seen in the first five conditions of experiment 1, as human judgments are reliably non-zero while $\Delta P = 0$. And as the base-rate of the effect decreases, so do strength judgments.

Advocates for both ΔP and causal power have attempted to explain the frequency illusion by focusing on various aspects of the experimental method. For instance, Shanks (1995) hypothesizes that the illusion may be due to measurement of judgments before they reach equilibrium. He shows that in non-contingent conditions, and with a certain choices of parameters, Rescorla-Wagner learning will initially assign nonzero association strength before eventually reaching an equilibrium of 0.

Proponents of causal power argue that the frequency illusion, as well as other deviations from power, are experimental artifacts (Buehner & Cheng, 1997; Buehner et al., 2003; Liljeholm & Cheng, 2009). Accordingly, they contend that proper revisions to the experimental method will minimize the influence of so-called extraneous factors, so as to better isolate the causal judgment process. Memory biases and ambiguity of the casual strength question are identified as two potentially important extraneous factors. To minimize the influence of memory effects, Buehner et al. (2003) conducted several of their experiments using the list format. Supporting their hypothesis, results from these conditions were in better accord with causal power predictions.

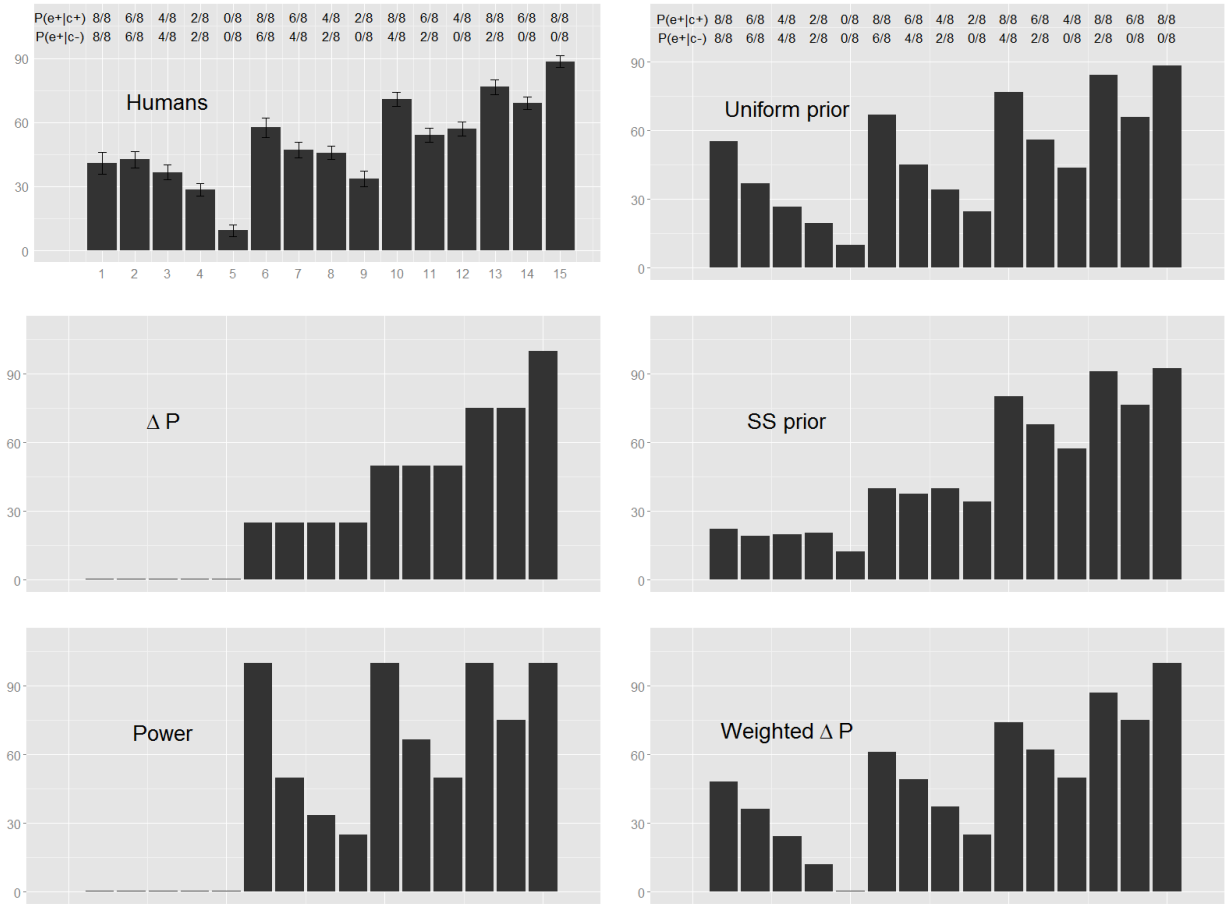


Figure 1.2. Predictions of causal learning models compared to human judgments from Buehner & Cheng (1997, Experiment 1B). Numbers at the top show stimulus contingencies. Error bars indicate one standard error.

The wording of the causal probe question has also been a target of revision. Recall that the standard wording asks how strongly the cause produces the effect. Buehner et al. (2003) identify several ways in which this standard question may be deficient. They argue that the "how strongly" formulation is ambiguous and could imply either (a) the current learning context or (b) a counterfactual context in which "there are no alternative causes of like kind," (Buehner et al., 2003, p. 1126). Buehner et al. (2003) go on to assert that the first interpretation supports a ΔP judgment while the second interpretation is consistent with causal power.

Another potential problem is that ambiguity in the causal question could cause participants to conflate causal strength judgments with confidence judgments. The concept of a virtual sample size is used to explain how confidence will vary across conditions in which causal power is fixed (Liljeholm & Cheng, 2009). Following Liljeholm & Cheng's example, suppose there is a particular

mineral in an allergy medicine that causes headaches as a side effect. Now consider two hypothetical scenarios:

> **Scenario A:** Headaches occur on 15 out of 20 trials before taking the mineral (the c^- trials) and 20 out of 20 trials after the mineral is given (the c^+ trials).
>
> **Scenario B:** Headaches occur on 5 out of 20 trials before taking the mineral (the c^- trials) and 20 out of 20 trials after the mineral is given (the c^+ trials).

If one assumes no confounding—that all alternative causes are independent across scenarios—then power-PC theory may be applied. Causal power is equal to 1 in both scenarios. In contrast, ΔP differs with $\Delta P = 1/4$ for Scenario A and $\Delta P = 3/4$ for Scenario B. Sample sizes are equal for both scenarios, with $N(.,c^+) = N(.,c^-) = 20$ and $N(.,.) = 40$. Yet Scenario B would seem to give stronger evidence that causal power equals 1.

Liljeholm & Cheng (2009) define virtual sample size as, "the estimated number of trials on which the production of headache can be unambiguously evaluated," (p. 159). The virtual sample size for Scenario A is 5, since these are the only trials in which the cause can demonstrate its efficacy. And for Scenario B the virtual sample size is 15.

The concept of virtual sample size is used in the conflation hypothesis, a two-part explanation for why judgments deviate from causal power (Buehner & Cheng, 1997; Buehner et al., 2003; Liljeholm & Cheng, 2009). The conflation hypothesis posits that: 1) confidence in causal strength judgment is an increasing function of virtual sample size and 2) measured causal strength and confidence judgments are conflated due to ambiguities in the experimenter's question.

Suppose causal power is fixed across several generative experimental conditions while ΔP varies. Then the conflation hypothesis predicts that deviations from causal power will track the magnitude of ΔP. In particular, as ΔP decreases the virtual sample size decreases, and so people will be less confident in their causal judgments.

If the conflation hypothesis is correct, and confidence infiltrates strength judgments, then careful measurement is necessary to disassociate the two. A number of studies have argued that a counterfactual or suppositional question wording better isolates strength judgments. In the revised questions participants are asked to imagine a baseline in which the effect never occurs, and then

predict the frequency of the effect once the cause is introduced. An example of the revised format from Liljeholm & Cheng (2009) is:

- Suppose that there are 100 people who do not have headaches.
- If this mineral was given to these 100 people, how many of them would have a headache?

A number of researchers using this type of format have found judgments that better conform to causal power predictions (Buehner & Cheng, 1997; Buehner et al., 2003; Collins & Shanks, 2006; Liljeholm & Cheng, 2009).

Unsurprisingly, critics of causal power have not been persuaded by the conflation hypothesis. Perales and Shanks (2003) use a different paradigm to test the conflation hypothesis by having participants make causal strength judgments at different levels of confidence. In "high confidence" conditions, participants were instructed to continue studying butterfly records until they were 100% confident with their ratings. In these high confidence conditions, contingency was found to still influence judgment when causal power was held constant. Consequently, Perales and Shanks (2003) reject the conflation hypothesis.

There is also debate concerning the adequacy of the counterfactual question. Perales and Shanks (2008) speculate that the complex wording may lead to a non-normative interpretation by participants. Specifically, participants may believe they are being asked to imagine that the base rate $P(e^+|c^-)$ is equal to 0 and then to estimate the conditional probability $P(e^+|c^+)$. Results from Perales and Shanks (2008) were consistent with this hypothesis. The authors tested several versions of the counterfactual question in conjunction with both sequential and list presentation of the learning trials. For the counterfactual question they found that judgments were highly influenced by $P(e^+|c^+)$ while the base rate $P(e^+|c^-)$ was largely neglected. The standard causal probe, in contrast, produced judgments that were a function of both conditional probabilities $P(e^+|c^+)$ and $P(e^+|c^-)$. Chapter 5 returns to the discussion of the causal probe question and demonstrates conceptual problems with the counterfactual or suppositional wording.

1.6.2 Empirical performance of algorithmic models

Perales and Shanks (2007, 2008) have found the EI rule to provide a good description of causal judgments. The weighted ΔP rule gives similar predictions to the EI rule, though Chapter 3 will more closely compare these two models. For now, it suffices to note that the weighted ΔP rule has good empirical fit to the Buehner & Cheng (1997) data, capturing the major qualitative trends including the frequency illusion. This is can be seen in the bottom right panel of Figure 1.2. The weighted ΔP rule also gives similar predictions to one of the two Bayesian models discussed in the next section.

1.6.3 Empirical performance of Bayesian models

We have seen that the causal power model has empirical shortcomings and attempts to resolve them through changes to the experimental design have had mixed success. Another response to the problematic findings has been the development of Bayesian models, which were described above. The top right panel in Figure 1.2 shows predictions from a Bayesian model of causal power with a joint uniform prior. One can see that the uniform Bayesian model does an excellent job capturing the major qualitative trends, including the frequency illusion. The model also nicely fits data from the Lober and Shanks (2000) experiments discussed above (Griffiths & Tenenbaum, 2005).

Yet the machinery of Bayesian inference alone is not sufficient to produce a satisfactory model. Figure 1.2 also shows predictions from the SS prior model with Lu et al.'s (2007, 2008) preferred $\alpha = 5$. One can see that it misses the qualitative pattern of the frequency illusion. The SS prior predicts nonzero judgments, though they do not decrease with a decreasing base rate of the effect. Of course, this result may not seem troublesome for those who are dubious about the reality of the frequency illusion. Yet more problematic are the predictions for conditions 6 through 9 in which ΔP is fixed at ¼ while power systematically decreases, as do human judgments. Again, the SS prior model misses this ordinal prediction while the uniform prior model does predict this trend.

There has also been some work comparing Bayesian models using different generating functions. Lu et al. (2007, 2008) investigated Bayesian models using the linear and the causal power generating functions in conjunction with both SS and uniform priors. They found that

models using the power generating functions performed much better than those with a linear function. Specifically, the Bayesian models of causal power had a higher correlation and a lower root mean-squared error when evaluated against human judgments. In addition, this empirical advantage held when using either uniform or SS priors.

1.6.4 Summary of findings

On reviewing the totality of the arguments and evidence, most researchers would likely agree that causal power is superior to ΔP as a normative theory of causal judgment. Yet most would also agree that neither model is empirically adequate. Problematic findings for causal power include the frequency illusion and judgments that reliably vary with ΔP when power is held fixed. There have been two general strategies for dealing with these problems. The first, as exemplified by Buehner et al. (2003), is to amend experimental design in order to minimize the influence of so-called extraneous factors. The second strategy is to embed causal power within a framework of Bayesian inference while leaving experimental method alone. Both strategies buy some success, though each at its own cost.

One can argue that some of the above changes to experimental methodology have incurred a cost in ecological validity. Greater adherence to causal power was found with a list format, yet it seems unlikely that everyday causal judgment occurs over an array of observations. With regards to the causal question, there is a risk that the more technical language moves people away from making a natural assessment. Indeed, Perales and Shanks hold this view in stating,

> What does seem obvious is that accurate power estimation requires a compatible probability-based presentation format and strong guidance by the experimenter. This seems to contradict the idea that computing causal power is intuitive and based on the existence of a module for detection of causality. (Perales & Shanks, 2008, p. 1493)

In experimental work it is necessary to strip away some normally occurring factors in order to isolate the process of interest. The constant challenge is to distinguish between the factors that only add noise or bias and the factors that are necessary for the psychological process to become manifest. A researcher may believe that they are only removing the noise or bias factors when in

fact they are also eliminating necessary factors. When this is the case, behavior in the amended design may be even more artifact than what was observed in the original.

The application of Bayesian inference also has a price. Most obviously, an additional layer of complexity is introduced. Much less obvious are the consequences and their severity. This issue will be taken up in detail in the coming chapters. It is first instructive to better understand what the Bayesian models are providing beyond a better fit to human data. The basic mechanics of the Bayesian approach were shown above, but there was no specific account for why it is useful for the problem of causal inference.

The next chapter illustrates the value of Bayesian inference with a simple example. Lessons from the example are then applied to the problem of causal inference, which should make clear what the Bayesian model is achieving. It will then be apparent why a rational agent should prefer a Bayesian model over the causal power MLE. Nonetheless, controversy surrounding the Bayesian approach in cognitive science remains. Chapter 2 also reviews this debate and explore its implications for field of causal learning.

Chapter 2.

A closer look at Bayes

2.1 Why Bayes?

Suppose you just moved to a new city and you hear that the incumbent mayor Nick Nickerson is running for re-election against challenger Patti Pattison. You are completely ignorant of the local politics, but from previous experience you know incumbent mayors win about 60% of the vote. However, there is considerable uncertainty around this 60% average, with both higher and lower values being plausible. Over the course of one day, you talk to three self-identified voters and discover that one supports the incumbent Nickerson while the other two support Pattison (you are quite gregarious, so you assume these three are close to a random sample).

What is your best guess for the percentage who will vote Nickerson? One option is to guess 33%, the sample proportion of the voters you talked to. The sample proportion gives the maximum likelihood estimate. Another option is to combine your prior knowledge with the sample information. As was shown in Chapter 1, Bayes rule gives one method to perform this combination.

To be more specific, we can describe our prior using a beta(3,2) distribution (Figure 2.1, blue curve). This gives a prior expectation of 0.6 (vertical blue line) while the prior mode is 2/3. The prior distribution around its mean and mode is relatively flat, reflecting belief that does not change much for the values over this range. In other words, there is uncertainty in prior belief since many values are almost equally probable.

The yellow curve plots the likelihood function for our sample (the vertical scale has been transformed to be commensurate with the prior). The vertical yellow line is the sample proportion of 1/3. The sample proportion occurs at the maximum of the yellow curve and so it gives the MLE.

Finally, the green curve shows the posterior distribution, which is found by the combination of the prior and the likelihood using Bayes rule. The posterior expectation is often used as a Bayesian point estimator. The green vertical line in the figure gives the posterior expectation, which is equal to 1/2. The posterior expectation will always be intermediate of the prior expectation and the MLE. In this example, the posterior mean is closer to the prior mean than it is to the MLE.

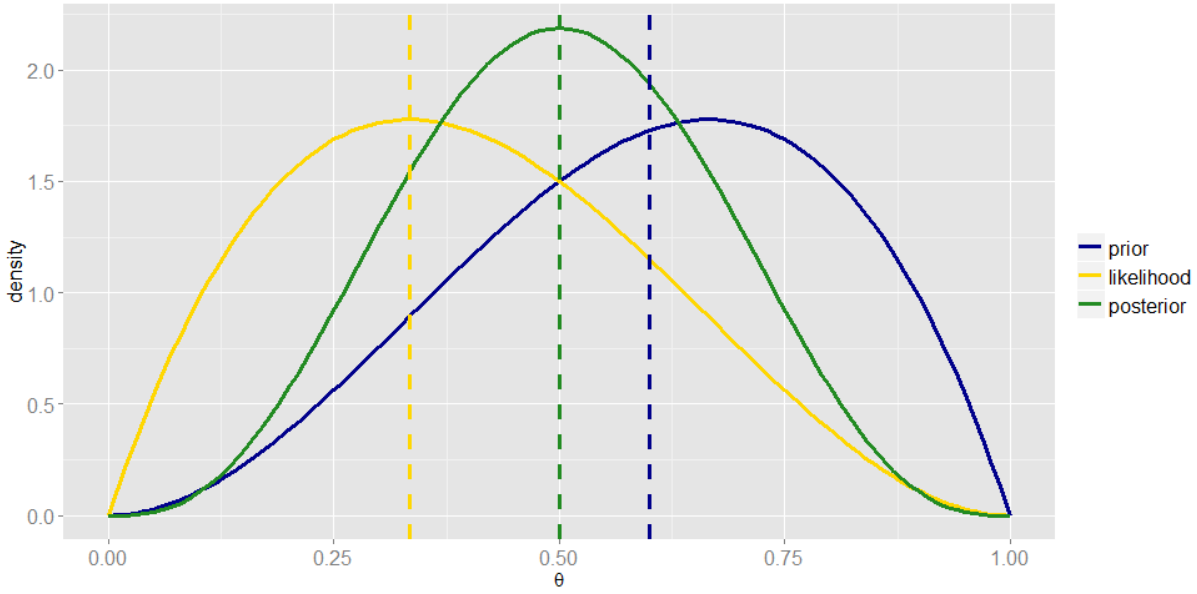


Figure 2.1. Plots of the prior (blue), likelihood (yellow) and posterior distribution (green). Correspondingly colored vertical dotted lines mark the prior expectation, MLE, and posterior expectation.

Now what should be used as the point-estimate for our election prediction? The MLE, the posterior mean, or something else? One way to answer this question is to compare the sampling properties of the candidate estimators, which is just the average performance of the estimator across many random samples. For instance, we may be interested if, on average, the estimator hits the target it is estimating. An estimator is unbiased if, across many repeated samples, its average value is equal to the population value.

Returning to the election example, suppose that θ_t is the true proportion who will vote for Nickerson. Denote the maximum-likelihood estimator as $\hat{\theta}_{\text{MLE}}$ and the posterior mean estimator as $\hat{\theta}_{\text{post}}$. It can be shown that the MLE is unbiased, or formally, $E\left[\hat{\theta}_{\text{MLE}} \middle| \theta = \theta_t\right] = \theta_t$. Another important property of an estimator is how close it typically is to the true value. One common way to measure this is with the mean-squared error or MSE. An estimator's MSE can be decomposed into a variance and a bias component (Hoff, 2009). For a given estimator $\hat{\theta}$,

$$\text{MSE}\left[\hat{\theta} \middle| \theta = \theta_t\right] = \text{Var}\left[\hat{\theta} \middle| \theta = \theta_t\right] + \text{Bias}^2\left[\hat{\theta} \middle| \theta = \theta_t\right] \tag{2.1}$$

There is usually a trade-off between the variance and bias of estimators, especially when sample information is limited and the true data generating function is complex (Hastie, Tibshirani,

& Friedman, 2009). Unbiased estimators will often be high in variance. For instance, if we had polled a single person, then the $\hat{\theta}_{\text{MLE}}$ would equal 0 or 1, both seemingly implausible estimates given what we know about elections in the U.S.

The posterior mean $\hat{\theta}_{\text{post}}$ will be biased towards the prior expectation so that $E\left[\hat{\theta}_{\text{post}} \middle| \theta = \theta_t\right] \neq \theta_t$ (it will be unbiased only if the prior expectation exactly equals the population value θ_t). But so long as the judge has some minimal information about the population sampled from, then the reduction in variance will be larger than the increase in squared bias. The upshot is that the Bayesian estimate $\hat{\theta}_{\text{post}}$ will have a lower MSE than will the sample estimate $\hat{\theta}_{\text{MLE}}$.

What happens to the Bayesian estimate as sample information improves? In the simple election example the posterior expectation $\hat{\theta}_{\text{post}}$ can be expressed as a weighted average of $\hat{\theta}_{\text{MLE}}$ and the prior expectation θ_{prior}:

$$\hat{\theta}_{\text{post}} = (1 - \beta) \times \theta_{\text{prior}} + \beta \times \hat{\theta}_{\text{MLE}} \tag{2.2}$$

The weight β is a function of the sample size *N*. As *N* increases so does β and more weight is given to $\hat{\theta}_{\text{MLE}}$. This seems sensible since with a larger *N* the sample estimate $\hat{\theta}_{\text{MLE}}$ is more reliable. Equivalently, the standard error of $\hat{\theta}_{\text{MLE}}$ becomes smaller as *N* becomes large, and this corresponds to high curvature and a peaked likelihood function at $\hat{\theta}_{\text{MLE}}$.

Returning to the election example, suppose instead that you see the local paper has conducted a poll of likely voters for the upcoming election. Out of 300 polled, 100 support Nickerson and 200 support Pattison. The beta(3,2) prior, the likelihood (again scaled), and the posterior distribution are all shown in Figure 2.2.

With the larger sample size, the $\hat{\theta}_{\text{MLE}}$ estimate is much more precise, which is reflected by the peaked likelihood. Now the posterior distribution almost overlaps the likelihood and the expectations are nearly equal with $\hat{\theta}_{\text{post}} \approx \hat{\theta}_{\text{MLE}}$. When the evidence is strong, belief mainly depends on the sample information.

Thus, a Bayesian estimate can be a good pragmatic choice insofar as it typically has a lower average squared distance to the truth than the maximum likelihood estimate. However, one often sees much stronger statements for the virtues of Bayesian inference, sometimes describing it as

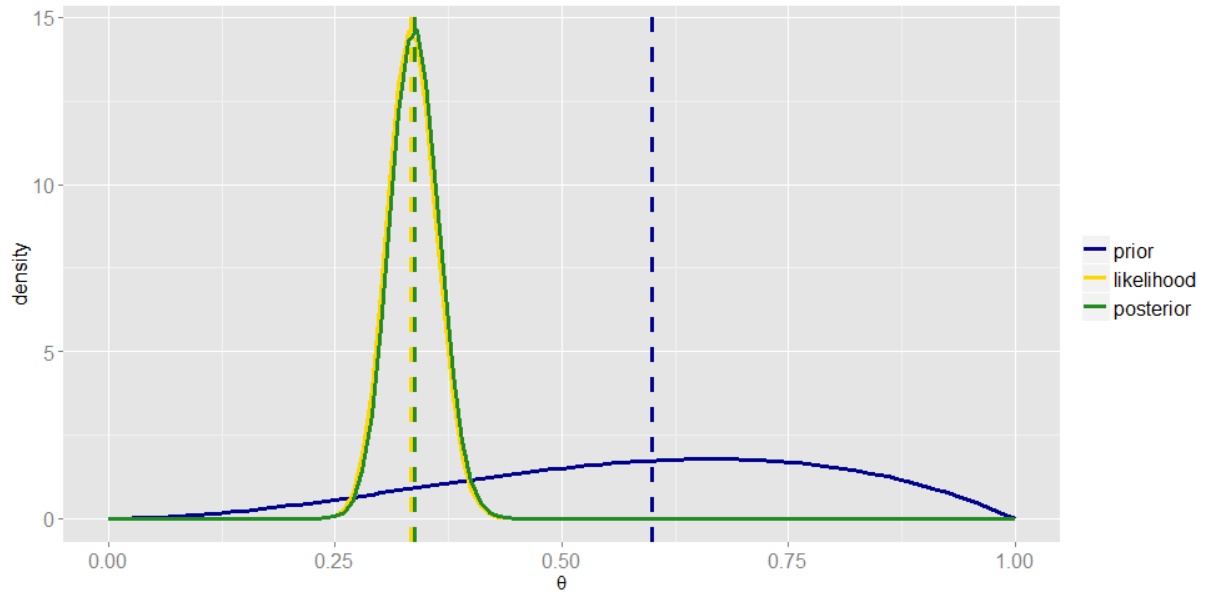


Figure 2.2. Plots of the prior (blue), likelihood (yellow) and posterior distribution (green). Correspondingly colored vertical dotted lines mark the prior expectation, MLE, and posterior expectation.

"optimal" or "normative". These descriptors reference results from decision theory. On this approach, one specifies a "loss function" that returns a penalty whenever the estimate does not equal the true value. Specifically, suppose the estimator $\hat{\theta}$ is used to estimate the true value θ while loss is given by the function $L[\theta, \hat{\theta}]$. Then a standard assumption is that $L[\theta, \hat{\theta}] \geq 0$ for all $\theta, \hat{\theta}$, and that $L[\theta, \theta] = 0$, so loss is zero only when the estimate equals the true value (Lehmann, Casella, & Fienberg, 1998). If a judge has squared error loss over their posterior, or $L[\theta, \hat{\theta}] = (\theta - \hat{\theta})^2$, then the posterior mean will minimize expected loss (Jaynes, 2003). It is in this sense, then, that the Bayesian estimate is optimal. Squared error has traditionally been a popular choice for the loss function, but there are good reasons for other choices. For instance, if loss is given by the magnitude of the error, or $L[\theta, \hat{\theta}] = |\theta - \hat{\theta}|$, then the posterior median will minimize expected loss (Jaynes, 2003).

2.2 Bayesian causal power revisited

Hence, the posterior expectation is a combination of the prior expectation and the MLE, with the weight given to each determined by how much information is in the sample. These ideas also inform Bayesian models of causal power. In the simple election example, the strength of the

sample evidence was determined only by the number of observations. The relationship is a bit more complex for causal power. In this case, sample information depends on the number observations and, quite crucially, the base rate of the effect. The concept of virtual sample size provides intuition for the role of the base rate.

Recall that virtual sample size is the number of trials on which the cause can be unambiguously evaluated. It is easy to see that, for a fixed level of causal power, virtual sample size is determined by the base rate of the effect. Return to the example from Section 1.6.1 with Scenarios A and B in which power is fixed at 1. Scenario A has a high base rate of $P(e^+|c^-) = 0.75$, which corresponds to a virtual sample size of 5. The low base rate Scenario B of $P(e^+|c^-) = 0.25$ gives a virtual sample size of 15. People's intuitions seem to reflect that Scenario A gives less reliable information, as Liljeholm & Cheng (2009) found that higher confidence ratings were recorded for conditions with a larger virtual sample size.

The influence of the base-rate on the reliability can be characterized precisely. This is most clearly demonstrated using the weights notation from common-effect causal graphs. Recall that w_0 gives the background causal strength for *B* and w_1 is the causal strength of the candidate cause *C*. Finally, w_T is the "total" causal strength for the conjunction of the causes *C* and *B*. The bottom row of the contingency table (Table 1.1) is used to estimate $\widehat{w}_0 = \hat{P}(e^+|c^-)$ while the top row is used to estimate $\widehat{w}_T = \hat{P}(e^+|c^+)$. Hence, $\widehat{w}_0$ and $\widehat{w}_T$ are both random quantities with $\widehat{w}_0 \sim \frac{1}{N}\mathrm{Bin}(w_0, N)$ and $\widehat{w}_T \sim \frac{1}{N}\mathrm{Bin}(w_T, N)$ when there are *N* observations for each row. Griffiths and Tenenbaum (2005) show that the maximum likelihood estimator for causal power is:

$$\widehat{w}_1 = \frac{\widehat{w}_T - \widehat{w}_0}{1 - \widehat{w}_0} \tag{2.3}$$

In order to simplify the exposition, I will assume that background strength w_0 is known and fixed. The causal power MLE then becomes:

$$\widehat{w}_1 = \frac{\widehat{w}_T - w_0}{1 - w_0} \tag{2.4}$$

While the assumption of a fixed base rate will not describe most applications, all of the essential conclusions to follow will still hold for a random base rate $\widehat{w}_0$. However, the treatment of the random base rate case will be primarily relegated to Appendix B.

Now the goal is to understand the relationship between the base rate and the reliability of the causal power MLE. As above, the mean squared error is used to measure reliability. Since (2.4) is an unbiased estimator of causal power, the MSE will just equal to the variance:

$$\mathrm{Var}[\widehat{w}_1] = \frac{(1 - w_1)}{N} \times \left[\frac{w_0}{1 - w_0} + w_1\right] \tag{2.5}$$

The importance of the base-rate relative to the sample size is made clear from (2.5). As $w_0 \to 1$ the variance becomes arbitrarily large. If we fix N and w_1, then the variance increases in w_0 at the rate of $\frac{\partial}{\partial w_0}\mathrm{Var}[\widehat{w}_1] = \frac{1}{(1-w_0)^2} \times c(N, w_1)$, where $c(N, w_1)$ is a constant based on N and w_1. Since $0 \leq w_0 \leq 1$, this indicates a quadratic rate of change in w_0. Similarly, for fixed w_0 and w_1, the variance decreases in N at the rate of $\frac{\partial}{\partial N}\mathrm{Var}[\widehat{w}_1] = -\frac{1}{N^2} \times c(w_0, w_1)$, also a quadratic change. Thus, the base rate and sample size have commensurate influence on the MSE.

The derivation for (2.5) is in Appendix B. The random base rate case of (2.3) is a ratio of random variables, and in general it is not possible to find exact expressions for its expectation or variance. However, approximations can be found, which are also shown in Appendix B.

The causal power mean squared error in (2.5) is actually a bit of a simplification. It assumes that (2.4) is always applied, though no reasonable judge would use the (2.4) estimator for $\widehat{w}_T < \widehat{w}_0$, when ΔP is negative. Instead, they would probably apply a mixed strategy, using (2.4) for positive ΔP and either 0 or preventive power for negative ΔP. Within a generative context, such a strategy will reduce variance but add bias. The general conclusions of (2.5) still hold for this mixed estimator, as shown in Appendix B.

Recall the relationship between the variance of an estimate and the shape of the likelihood: A high variance estimate corresponds to a likelihood with little curvature, which in turn indicates that the data do not contain much information. Figure 2.3 gives a visual illustration of how the causal power likelihood depends on the base rate of the effect. The likelihood is plotted for six different levels of the base rate w_0 while the causal power MLE is fixed at 0.5 and the sample size is set to 10 across all of the panels.

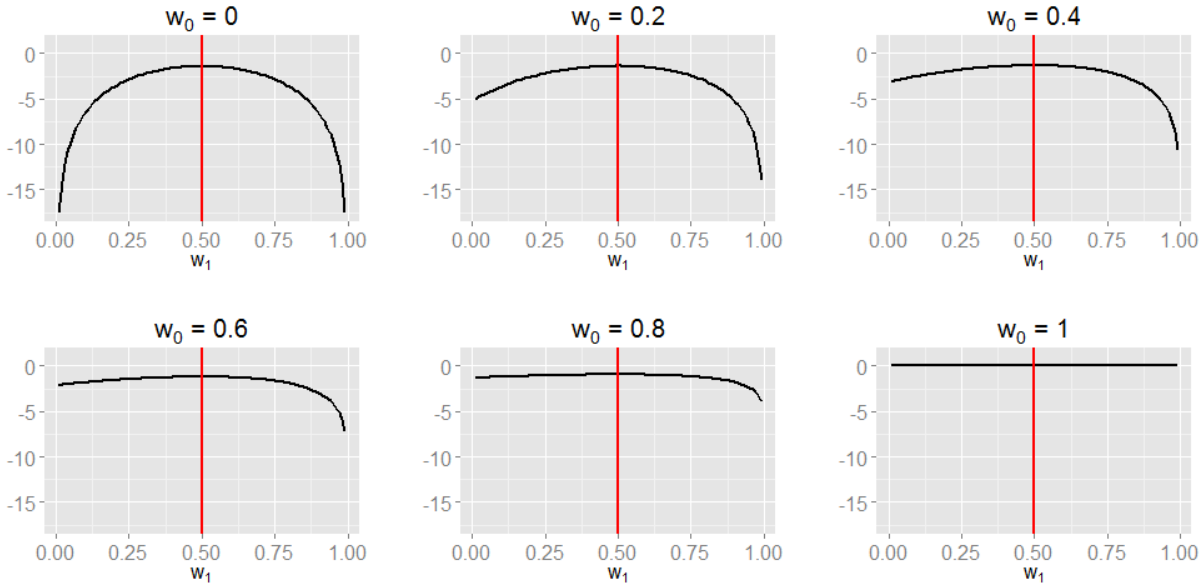


Figure 2.3. Causal power likelihood functions with MLE $\hat{w}_1 = 0.5$ (red line), sample size $N = 10$, and with each panel conditioned on a different base rate w_0. As the base rate w_0 increases, the likelihood flattens, which indicates that there is less information in the sample.

Figure 2.3 shows that the likelihood quickly flattens as the base rate of the effect increases. At $w_0 = 1$, the likelihood is perfectly horizontal, which shows that the sample contains no information.

This knowledge of the causal power likelihood can now be used to illuminate the mechanics of Bayesian models. The posterior judgment for causal power is a combination of prior belief and sample information. Denote the prior expectation by $w_{1.\text{prior}}$, the sample estimate as $\hat{w}_{1.\text{MLE}}$, and the posterior expectation as $\hat{w}_{1.\text{post}}$. First, consider the Bayesian causal power model with a joint uniform prior, so $w_{1.\text{prior}} = 0.5$. The extent that the posterior judgment reflects the prior expectation versus the sample estimate depends on how informative the data are. Equation (2.5) shows that, for a fixed level of causal power, the sample size N and the base rate of the effect w_0 determine the informativeness of the sample. As more observations accrue, the data give stronger evidence and $\hat{w}_{1.\text{post}}$ moves closer to $\hat{w}_{1.\text{MLE}}$. When the number of observations becomes arbitrarily large, $\hat{w}_{1.\text{post}}$ will eventually converge to $\hat{w}_{1.\text{MLE}}$.

The base rate of the effect has the opposite influence: as w_0 increases, there is less information in the data and $\hat{w}_{1.\text{post}}$ will be closer to $w_{1.\text{prior}}$. This property of Bayesian causal power explains why predictions should vary with ΔP. For a fixed level of causal power, a smaller ΔP corresponds

to a higher base rate of the effect w_0. Consequently, as ΔP becomes smaller, predictions should be more regressive to the prior expectation $w_{1.\text{prior}}$. This explanation also covers the frequency illusion. All conditions with $\Delta P = 0$ give a data estimate of $\widehat{w}_{1.\text{MLE}} = 0$ (so long as $w_0 < 1$). When $w_0 = 0$, the data are most informative and so $\widehat{w}_{1.\text{post}}$ should be closest to $w_{1.\text{MLE}} = 0$. And as w_0 increases, $\widehat{w}_{1.\text{post}}$ will become more regressive to the prior expectation, so predicted judgments will increase away from $\widehat{w}_{1.\text{MLE}} = 0$ and towards $w_{1.\text{prior}}$. Finally, at $w_0 = 1$, there is no information in the sample and the posterior expectation should just equal the prior expectation. Thus, we see how a Bayesian causal power model with joint uniform priors can explain key patterns in human judgment.

An essential assumption for the preceding argument is that background and candidate strengths are independent in the prior. This implies that $w_{1.\text{prior}}$ does not depend on the level of w_0. Consequently, prior expected strength remains constant at $w_{1.\text{prior}} = 0.5$ across varying levels of the base rate w_0. The next section examines the implications of the dependence between w_0 and w_1 in the SS prior model.

2.3 Trouble with the SS prior model

The SS prior model does a poor job describing the frequency illusion (Figure 1.2, top-left panel, conditions 1 through 5), and it even fails to capture some ordinal predictions of causal power (Figure 1.2, top-left panel, conditions 6 through 9). Why is this? Recall that SS priors encode beliefs of dependence between strengths w_0 and w_1. Specifically, for a generative SS prior there is competition in belief for the two causal strengths with peaks in the $f(w_0, w_1)$ joint density over (1,0) and (0,1). When there are two generative causes, and one has a stronger association with the effect in the data, the SS prior will enhance this difference by reducing the judged strength of the weaker cause. Figure 3 of Lu et al. (2008) nicely shows the influence of the SS prior relative to the joint uniform prior.

Competition in causal strengths explains important empirical difficulties of the SS prior model. With an SS prior, an increase in the base rate w_0 influences the posterior estimate in two ways. As before, there is less sample information so the estimate becomes more regressive to the prior expectation. But in addition, the prior expectation $w_{1.\text{prior}}$ decreases as w_0 increases because of the competition between w_0 and w_1. In the case of the frequency illusion, these two influences

are opposing since $\widehat{w}_{1.\mathrm{MLE}} = 0$. As w_0 increases, the prior has a larger influence relative to $\widehat{w}_{1.\mathrm{MLE}} = 0$. At the same time, the prior expectation is decreasing. As can be seen in Figure 1.2, these two influences roughly offset, giving predictions that are fluctuating and relatively flat.

The same issue occurs in conditions 6 through 9 from Figure 1.2. The causal power MLEs for these conditions are, respectively, 1, 0.75, 0.5 and 0.25 while the base rates are 0.75, 0.5, 0.25 and 0. So the causal power MLE is positively associated with the base rate in these conditions. As before, a strong base rate produces greater regression to the prior. But now, competition in the SS prior roughly negates the sample evidence. The result is a fluctuating/flat prediction across conditions, contrary to the monotonic trend observed in human judgments.

To summarize, the dependence assumed by the Bayesian SS prior model gives predictions that are incompatible with human judgments. Lu et al. (2008) did not account for this empirical shortcoming, and I am unaware of any subsequent authors who have. In contrast, a Bayesian model with a joint prior that is independent in w_0 and w_1 can explain why human causal strength judgments should vary both with changes in causal power and with changes in ΔP.

2.4 Artifact or rational inference?

Two competing explanations have been presented for why judgments systematically deviate from causal power. On one view, the deviations are experimental artifact to be eliminated. The Bayesian framework, in contrast, interprets the deviations as sound judgments in the face of sample uncertainty. Reliability is central to both accounts, though it takes a very different explanatory role in each. Reliability in the conflation hypothesis, as expressed through confidence, is viewed as a confound to be removed. Whereas on the Bayesian view, reliability is an input to optimal inference, used to balance prior knowledge with sample information.

Might there be some way to reconcile these two strategies? Certain elements appear consistent across the two accounts. For instance, the list format, by reducing memory demands, may effectively increase the sample size. This would make Buehner et al.'s (2003) finding that the list format better elicits judgments closer to $\widehat{w}_{1.\mathrm{MLE}}$ consistent with predictions of the Bayesian model. However, so long as sample information is imperfect, there will be a fundamental disagreement between approaches. The conflation hypothesis holds that the true psychological assessment is $\widehat{w}_{1.\mathrm{MLE}}$ and for the Bayesian model it is $\widehat{w}_{1.\mathrm{post}}$. The disagreement will be most pronounced when

the sample size is small or the base rate of the effect is high. As an illustration, imagine a condition in which participants are presented learning data in a list format with 99/100 successes in the c^- condition and 100/100 successes in the c^+ condition. The causal power MLE in this example is $\widehat{w}_{1.\mathrm{MLE}} = 1$ while the joint uniform prior Bayesian model gives $\widehat{w}_{1.\mathrm{post}} \approx 0.6$.

The Bayesian explanation appears better justified on two counts. First, Bayesian inference will be optimal when model assumptions are met. But even if the model is only approximately true, the Bayesian prediction will most likely have a lower mean-squared error than the maximum likelihood estimator. On the second count, the Bayesian model describes human judgments from the original experimental design with the sequential format and standard causal probe. Most agree that this format has higher ecological validity, and so judgments from these experiments will more closely correspond to natural assessments.

It would seem, then, that the Bayesian approach is best for the problem of elemental causal induction. On this approach rational Bayesian inference is combined with the normative power PC model. The only complication is to use an acceptable prior. As we've seen, the joint uniform density does quite well and it can also be justified a priori. Alternatively, one can search for the appropriate prior using human judgments, as is done by Yeung and Griffiths (2015).

The combination of Bayesian inference with the rational approach has become increasingly widespread. Rational Bayesian models have been proposed for nearly every major phenomenon in cognitive science (see Eberhardt & Danks (2011) or Endress (2013) for long lists of references across multiple domains). But more recently, there has been criticism of the rational Bayesian program. The next section reviews these criticisms and considers how they apply to Bayesian models of causal learning. The discussion is then used to guide the task of integrating rational and algorithmic approaches, which is taken up in Chapter 3.

2.5 The rational Bayes debate

To begin, revisit Anderson's (1990) conception of the rational approach described in the introduction. First, an agent's goals and its environment are precisely characterized, and then the optimal solution is derived using this characterization. Next, model predictions are compared to human judgments and the model is revised in order to accommodate discrepancies. The process is iterated, and it is assumed that the testing-revision cycles will lead towards a true description of

goals and the environment. Accordingly, the models should move closer to an optimality-based explanation of behavior. There have been a number of criticisms of the rational Bayesian approach, several of which are reviewed below.

2.5.1 Criticism 1: Optimality claims are not justified

A number of cognitive science papers have claimed to show that various types of behavior is optimal. Yet critics question whether such conclusions are defensible. Danks (2008) asserts that the standard rational analysis is missing a crucial component required to justify optimality claims. Specifically, it is necessary to show how the cognitive process converged to the specific proposed solution. This would typically involve an evolution by natural selection story, or an individual development based account. The key issue is that just because some solution constitutes an optimum, it does not imply that the solution was available to be selected by the optimizing mechanism.

Sloman and Fernbach (2008) explore the normative implications of model revision. They argue that the chief burden of a normative analysis is to show that a particular normative model is, in fact, the optimal way to perform a task. On this view, normative analysis is by definition an a priori endeavor. Thus, a model loses any claim to normativity once it is amended to better fit the data. After revision the model can only be justified as descriptive.

The problem of revision becomes clear when one reflects on why behavior may deviate from the normative model prescriptions. Differences may be due to either: 1) genuine non-normative behavior, 2) unfounded assumptions of the analysis, or both 1) and 2). To the extent that the initial normative analysis is convincing, one has more confidence that the observed behavior is in fact non-normative. Yet for a given normative model and dataset, it is impossible to know with certainty where the problem lies. One response to this problem has been to double down on the assumption that behavior is optimal. Oaksford and Chater (2007) best exemplify this view, arguing that the relevant normative theory will be determined, in part, by empirical human data. Taken to the extreme, observed deviations will only be evidence for model misspecification. Though clearly, making model revisions to fit the data and then labeling the product rational is circular (Sloman & Fernbach, 2008).

2.5.2 Criticism 2: Bayesian models are underconstrained

Several critics believe that the rational Bayesian approach has been falsely advertised as a more principled and constrained method for model building (Bowers & Davis, 2012a; Jones & Love, 2011). If anything, the critics argue, Bayesian models have been particularly susceptible to strong fitting of the data.

A principled approach to Bayesian model construction is to ground the model in objective measurements of agent and environmental attributes. However, as Jones and Love (2011) observe, this is rarely ever done in practice. In their estimation, the majority of rational analyses do not include measurements from the actual environments of interest. Likewise, subjects' prior beliefs are almost never measured independently. Instead, researchers have demonstrated enormous leeway in the choice of prior, generative model and utility function. Jones and Love (2011) argue this indeterminacy allows for highly flexible models that can fit most any pattern of data. At the same time, the model building process is obscured by multiple potential loci of revision.

Bowers and Davis (2012a) level many similar criticisms as Jones and Love, again emphasizing the extreme flexibility of Bayesian models. In fact, they make the stronger claim that this flexibility results in models that are unfalsifiable. Bowers and Davis (2012a) supplement their critique with many examples, drawing on rational analyses of speed perception, word identification and higher level cognition. In each of their examples, they show that empirically driven revisions are not small tweaks, but instead are critical to model success. Endress (2013) also attacks the flexibility of rational models, claiming that the success of Frank and Tenenbaum's (2011) Bayesian model of rule-learning hinges crucially on what, in his view, is an ad hoc assumption.

Marcus and David (2013) look at the performance of rational models over eight different cognitive science research domains. They argue that so-called optimal performance often depends on post hoc choices for the priors and loss-functions (though see Goodman, Frank, Griffiths, Tenenbaum, Battaglia, and Hamrick (2015) for a response). In addition, Marcus and David (2013) show that a large proportion of studies in their survey suggest that humans behave non-optimally. Of course, whether behavior is optimal is largely in the eye of the beholder, but this difficulty only serves to underscore the point.

A virtue often attributed to rational analysis is that it can serve to constrain lower level psychological theories. Yet if the above criticisms are correct, Bayesian rational models are essentially just a re-description of the dataset. Accordingly, the constraints from rational analysis will add nothing beyond the constraints implied by the dataset (Bowers & Davis, 2012a). Bayesian models may still be useful as descriptive tools (Danks, 2008; Eberhardt & Danks, 2011). However, this descriptive job might also be achieved by a class of simpler models (Eberhardt & Danks, 2011).

Eberhardt and Danks (2011) present a novel critique in observing that almost all Bayesian models of cognition fail to specify how participants make choices based on their posterior distributions. This is problematic, as some choice principle must be included to make predictions. A seemingly natural principle is the maximization of posterior utility, yet this does not appear to be compatible with empirical findings. Namely, it is typically found that the *aggregate* response pattern resembles the predicted *individual* posterior distribution (e.g see Schulz, Bonawitz & Griffiths (2007)). For example, suppose the hypothesis space consists of only two options, A and B, and that an agent has a posterior probability of 0.6 for A and 0.4 for B. If utilities are equal across options, then an agent should choose A to maximize posterior utility. Notwithstanding some measurement error, the aggregate prediction is that everyone should choose A. Instead, what is typically observed is that about 60% of people choose hypothesis A and 40% choose hypothesis B.

A potential explanation for the aggregate response pattern is found with the "probability matching" hypothesis. Probability matching refers to a choice procedure where individuals randomly sample from their posterior distribution in order to choose hypotheses (or they use some method that is equivalent to this procedure). Response patterns are consistent with the probability matching hypothesis. However, they clearly do not imply probability matching. For instance, the aggregate pattern could also result each individual using a heuristic strategy (Mozer, Pashler, & Homaei, 2008).

Thus, additional evidence is necessary to show that people do in fact probability match. Eberhardt and Danks (2011) also argue that some account needs to be given for why people would probability match instead of maximize posterior utility. Vul, Goodman, Griffiths, and Tenenbaum (2014) attempt to provide such an account by showing that taking a few or even one sample from the posterior will be optimal when posterior samples are costly. But this is precisely the type of

move that causes ire among Bayesian critics. Any behavior can be found to be Bayesian optimal by post hoc amendments to model assumptions, such as specifying a cost for obtaining posterior samples. This is why Eberhardt and Danks (2011) call for explanations that are established on independent grounds for why people might engage in probability matching behavior.

2.5.3 Criticism 3: Rational approach neglects lower levels of analysis

Another criticism concerns the explicit agnosticism of the rational approach with regards to underlying psychological mechanisms. To justify this omission, the rational analyst's common refrain is that human cognition is extremely flexible, and so it should be able to support the sophisticated inferences implied by rational models. Yet closer scrutiny demonstrates the inadequacy of such a blanket defense. Kwisthout, Wareham, and Rooij (2011) argue that many Bayesian models of cognition require calculations that are computationally intractable. More sobering, even approximations for some of these models have been shown to be intractable. The authors conclude that for Bayesian models to preserve credibility, they must also include feasible algorithmic accounts.

Some critics emphasize that neglecting lower levels of analysis has been detrimental to the overall psychological enterprise. Jones and Love (2011) argue that rational analysis alone cannot be a sufficient account since psychological mechanisms will, at best, only approximate the optimal solution. A key benefit of a mechanistic understanding is that it generates novel predictions in the form of characteristic deviations from optimality. Bowers and Davis (2012a) give one example of this type: McClelland, McNaughton and O'Reilly's (1995) prediction of two separate learning systems in the hippocampus and neocortex. This prediction was based on the property of parallel distributed processing, a mechanistic model of representation. Danks (2008) also argues that mechanistic models allow for greater generalization of a cognitive theory. In particular, he contends that mechanistic models are better suited to generate predictions for non-standard environments or for cognitive systems experiencing deficits.

Others question the precedence of the computational level under the rational approach. For instance, Endress (2013) asserts that algorithmic explanations should come first, with computational theories developed only after an understanding of mechanism is in place. Further, Endress (2013, 2014) argues that pure computational models prove to be an elusive target. Hidden, yet crucial implementational assumptions can often find their way into a rational analysis. Endress

(2013) points to at least five implementational assumptions in Frank and Tenenbaum's (2011) computational Bayesian model of rule-learning, which he argues are all crucial for model predictions.

2.5.4 Criticism 4: Bayesian models rarely compared to alternatives

A fourth major criticism of rational Bayesian models is that they are rarely compared to either Bayesian or non-Bayesian alternatives. Jones and Love (2011) observe that for most Bayesian models of cognition there are alternative generating functions that can also be given strong justification. Consequently, they consider it a serious oversight when a model is presented without deliberate consideration and discussion of competitors. Bowers and Davis (2012a) emphasize the need to compare Bayesian to non-Bayesian models of the same process. Specifically, they state:

> In sum, given the computational complexity of implementing Bayesian computation in the brain, we would argue that the onus is on theoretical Bayesians to show that models that do implement such computations can explain human performance better than non-Bayesian models. This, as far as we are aware, has never been demonstrated. (Bowers & Davis, 2012a, p. 402)

A number of researchers have applied non-Bayesian models to challenge previous findings that purportedly demonstrate optimal Bayesian inference. Mozer et al. (2008) argue that a simple heuristic model better explains the data of Griffiths and Tenenbaum (2006) when compared to the original Bayesian account. For causal structure learning, Fernbach and Sloman (2009) and Bes, Sloman, Lucas, and Raufaste (2012) both preferred non-Bayesian over Bayesian models. And in his critique of Frank and Tenenbaum's (2011) Bayesian model of rule-learning, Endress (2013) proposes that simple psychological mechanisms can better account for the data.

2.5.5 Defense of rational Bayes

Naturally, proponents of rational Bayesian models have mounted a vigorous defense to the above critiques. These counterarguments are briefly summarized below.

Bayesian advocates agree that a very high standard must be met to warrant a claim of optimality. But they disagree with the charges that 1) the goal of rational analysis is to show that

people are optimal or 2) optimality claims are commonplace in the literature (Frank, 2013; Griffiths, Chater, Norris, & Pouget, 2012). Indeed, Frank (2013) states that "The standards for such a claim are almost never met," (p420) while Griffiths et al. (2012) hold that "In rare cases Bayesian models are used to argue that people behave optimally on a specific task," (p416). Yet Bowers and Davis (2012b) remain unconvinced, accusing Bayesian proponents of moving the (optimal) goalposts.

In her defense of the Bayesian approach, Hahn (2014) describes how a model's normative status is a function of the underlying cognitive task. Specifically, the force of the normative claim depends on how plausibly the task may be construed as one of probabilistic inference. If it clearly involves inference, then it is safe to assume that the Bayesian model is normative. To take an extreme example, the task can just be a formal Bayesian problem. Tversky and Kahneman (1982) describe one such problem administered by Casscells, Schoenberger and Grayboys (1978): medical students were asked to make a prediction after given all the relevant quantities in order to apply Bayes rule. For this example, few would disagree that the Bayesian model is normative (though see Gigerenzer (1991) for one of the few). Some tasks, on the other hand, may be plausibly construed in multiple ways. When the task is reasonably construed without using inference, a Bayesian model will not be inherently more normative or rational. The task of category formation, for example, is not obviously one of probabilistic inference. Accordingly, Anderson's (1991b) Bayesian model should not be afforded special normative status relative to non-Bayesian models.

Thus, Bayesian models may or may not give the normative standard depending on task content. Regardless, advocates contend that the Bayesian approach provides a valuable framework to construct models of cognitive processing. This claim pertains to criticism 2, wherein Bayesian models are charged with being overly flexible and potentially unfalsifiable. Some assert that the criticism is misconstrued due to a conflation of *models* with *theoretical frameworks* (Frank, 2013; Griffiths, Chater, et al., 2012). A model is constructed to account for a specific phenomenon. Models are potentially falsifiable since their predictions are tested against measurements. Theoretical frameworks, on the other hand, provide a general perspective and a set of tools for creating models (Griffiths, Chater, et al., 2012). Theoretical frameworks are either productive of unproductive, though they are not directly falsifiable. Productive frameworks create models with novel predictions that are confirmed empirically. They unify previously disparate phenomena, and

open new domains of theoretical inquiry. In contrast, unproductive frameworks generate few novel insights and create models that consistently require ad hoc revisions.

The essential question, then, is whether the Bayesian approach can be a productive framework. Critics can still maintain that the problem of flexibility also applies to the framework level. Their examples may be taken to demonstrate that the Bayesian framework is not productive since it is used to construct idiosyncratic models that always fit the data while offering little additional insight. Yet it is important to consider whether these problems are due to inherent limitations of the Bayesian approach, or whether the problem pertains to how it is typically applied. Holyoak and Lu (2011) contend that the Bayesian framework has been essential to the study of causal learning. This claim is the focus of the next section. Hahn (2014) also uses a number of examples to illustrate the productivity of the Bayesian approach. She argues that the Bayesian framework has spurred theoretical progress in the study of human reasoning and argumentation, producing novel predictions as well as giving rise to new theoretical questions. She also observes that the charge of unfalsifiability is belied by the vigorous debate surrounding the empirical adequacy of various Bayesian models. Hahn concludes that the Bayesian framework may be subject to abuse, but this does not entail the wholesale rejection of the approach.

With regards to the third criticism from above, advocates of the Bayesian approach argue that computational and algorithmic models should not necessarily be viewed as competitors since they are proposed at different levels of analysis. Further, there is reason to believe that mechanistic accounts will typically provide a better fit than functional accounts (Griffiths, Chater, et al., 2012). This is because a correct mechanistic formulation will embody the constraints that a functional account ignores.

Many Bayesian researchers do recognize the importance of developing models at the algorithmic and physical levels. Towards this end, recent work shows how Bayesian inference can be implemented using various Monte Carlo methods such as with importance sampling (Abbott, Hamrick, & Griffiths, 2013; Shi, Griffiths, Feldman, & Sanborn, 2010), the particle filter (Levy, Reali, & Griffiths, 2009; Sanborn et al., 2010) and Markov chain Monte Carlo (Gershman, Vul, & Tenenbaum, 2012). The general strategy in these papers is to look to the computer science and statistics literature to find the best algorithms for approximating statistical inference. Such a strategy yields *rational process models* (Griffiths, Vul, & Sanborn, 2012; Griffiths, Lieder, & Goodman, 2015). These process models are rational in two different senses. The first is that, given

sufficient processing power, the models will become identical with the ideal Bayesian solution. And second, they are the best possible known methods for approximating Bayesian inference when computational resources are limited. On this approach, then, the optimality of rational analysis can be viewed as being pushed down to the mechanistic level (Griffiths, Vul, et al., 2012).

2.6 The case of causal learning

How do Bayesian models of causal learning fare in light of the above debate? Consider the task of elemental causal induction. With respect to Hahn's (2014) continuum, there is a strong case for the probabilistic construal of the task. Chapter 1 showed that causal strength can be interpreted as a special type of probability. On this interpretation, then, causal strength estimation is a type of probabilistic inference. So if it is possible to specify a normative model, the Bayesian approach seems well suited to the job.

Next, consider how Bayesian models of causal learning have been formulated. Proposed generative functions include the ΔP rule (the linear generating function) and causal power (the Noisy-OR and the Noisy-AND-NOT). Importantly, these functions were not chosen based on their fit to human data. However, they also are not grounded in measurements of the objective environment. Indeed, it is not at all obvious what the relevant objective environment is. Is it the typical causal environment from our evolutionary history? Or is it the environments encountered over our individual life histories?

The difficulty of specifying the objective environment has led to a different strategy. Generative functions have been extensively justified on the basis of a priori argument. Similarly, both the uniform and SS priors are given a priori justification. This approach fits with Sloman and Fernbach's (2008) strict view of normative analysis from above. Thus, it seems that proposed Bayesian models of causal learning are at least potentially normative. Accordingly, challenges against their normative status should be targeted at the justification of these models. This is precisely what Cheng (1997) does in her rejection of the ΔP rule.

The Bayesian approach to causal learning also redeems itself on the issue of model comparison. As Holyoak and Lu (2011) observe, various Bayesian models have been proposed with different priors (uniform versus SS priors) and generating functions (linear versus the Noisy-

OR and the Noisy-AND-NOT). Bayesian models have also been compared to a plethora of non-Bayesian models, many of which were described in Chapter 1.

No specific account has been proposed for the implementation of Bayesian models of causal learning. At first blush, this does not appear to be especially problematic. Rational process models from above could also be applied to estimate a posterior distribution for causal strengths. But how much confidence should be placed in the rational process strategy? It shows one way to bridge levels of analysis. Yet independent evidence is still required to show that such algorithms are actually employed. Such evidence is necessary to rule-out non-Bayesian heuristic models that can mimic the Bayesian solution in key respects.

Independent supporting evidence can be roughly categorized as *direct* or *indirect*. One form of direct evidence is behavioral data, which is the standard for most psychological research. Behavioral data from causal learning experiments could indicate that judgments are distinctly Bayesian. While non-Bayesian models can approximate Bayesian predictions, it should still be possible to distinguish them on close examination using careful data analysis and innovative experiments. A second source of direct evidence comes from neuroscience. For instance, patterns of neural activation might be consistent with what one would expect from a system that is performing Bayesian computations.

Does the direct evidence strongly support a Bayesian model of causal learning? Behavioral data does suggest Bayesian-like inference, though it seems far from implying that behavior is distinctly Bayesian. At present, Bayesian models are not clearly preferred to non-Bayesian models in their empirical performance. Recall, for example, in Figure 1.2 that the Bayesian uniform prior and the weighted ΔP model both capture the dominant trends in human judgment.

Turning to neuroscience, researchers have explored how Bayesian inference might be implemented within the nervous system. The *Bayesian coding hypothesis* posits that populations of neurons represent uncertainty using probability distributions, which can then be used for Bayesian computations (Knill & Pouget, 2004). Work by Ma, Beck, Latham, & Pouget (2006) provides one solution using linear combinations of groups of neurons. While this work gives a potential solution, it does not show that neurons actually perform these computations. When Bowers and Davis (2012a) evaluate the neuroscientific evidence, they conclude that it offers little to no support of Bayesian theories. Further, they note that their view is shared by Knill and Pouget (2004), two key advocates of the Bayesian coding hypothesis (Bowers & Davis, 2012b).

Now if the direct evidence is equivocal, there will be considerable uncertainty about whether to prefer Bayesian or non-Bayesian models. It becomes necessary to rely more heavily on indirect criteria, such as representational commitments, coherence, memory requirements, and so on. If two models are equivalent empirically, then the standard preference is for the simpler model. For instance, if one model makes fewer representational and memory demands, then, ceteris paribus, it will be preferred. Such evaluations will necessarily be coarse, but this does not mean that they are arbitrary or ungrounded. Obtaining additional direct evidence from new and better experiments is always the long-term goal, but often one must begin with more informal criteria.

In fact, empirical fit along with informal "simplicity" criteria is what typically guides the development of mechanistic models. On such criteria, the weighted ΔP model appears to make fewer representational commitments than Bayesian models while delivering similar empirical performance. For these reasons, weighted ΔP merits additional consideration as a viable model of causal judgment. This is the focus of Chapter 3, which presents a novel computational treatment of weighted ΔP and shows that it can be naturally interpreted as an estimator of causal power. Chapter 4 then takes a closer look at the contrasting predictions of weighted ΔP versus Bayesian models. The conclusion will revisit the role of simplicity criteria in the selection of psychological models.

Chapter 3.

Bridging levels with weighted ΔP

3.1 Introduction

This chapter begins an analysis of weighted ΔP, a rule-based algorithmic model of causal judgment. Weighted ΔP is a non-Bayesian model, though it will be shown to share several desirable attributes with Bayesian inference. In this chapter I give an explicit computational account of weighted ΔP by interpreting it as an estimator of causal power. I then demonstrate how it may be implemented as a lower level process model. These findings, taken together, allow for novel derivations of Rescorla-Wagner models that attain a causal power equilibrium. Finally, I will explore weighted ΔP within the context of model uncertainty and show that it can be a reliable estimator when the underlying generative model is unknown.

3.2 Previous strategies to theory integration

Most researchers agree that cognitive science would benefit from tighter integration across the levels of analysis, as Jones and Love observe:

> The most accurate characterization of cognitive functioning is not likely to come from isolated considerations of what is rational or what is a likely mechanism. More promising is to look for synergy between the two, in the form of powerful rational principles that are well approximated by efficient and robust mechanisms. Such an approach would aid understanding not just of the principles behind the mechanisms…but also of how the mechanisms achieve and approximate those principles and how constraints at both levels combine to shape behavior. (Jones & Love, 2011, p. 186)

There are different views on the best way to bridge levels of analysis in cognitive science. Two general strategies can be distinguished by their choice of starting point:

Strategy 1: Begin with a normative computational model. Introduce resource constraints to explain deviations from normative predictions.

Strategy 2: Begin with a heuristic or algorithmic model. Use computational analysis to describe environments in which it performs well.

Variants of the first strategy have been proposed by Hahn (2014), Griffiths et al. (2012), and Sloman and Fernbach (2008), among others. The best known account of the second strategy is probably the fast and frugal heuristics program. On the fast and frugal approach, researchers begin with some heuristic model and then perform a computational analysis to determine the environments in which it succeeds or fails (Gigerenzer & Brighton, 2009; Todd & Gigerenzer, 2007). A heuristic is said to be *ecologically rational* if it is successful in the context or environment in which it is used. Another example of the second strategy is found with Ashby & Alfonso-Reese (1995), who begin with heuristic models of categorization and then determine what computational problems they appeared to solve.

In theory, each of the two strategies could produce the same final models. Yet in practice, the choice of initial model typically influences the outcome. Bridges travel only so far. Consequently, each strategy should produce models that inherit some of the strengths and weaknesses of the initial model. The first strategy will generate models with strong justification, though they may be less plausible as descriptions of the psychological process. Conversely, the second strategy will start with a psychologically plausible model, but a computational analysis committed to this model may be elusive or not particularly informative.

An example of the first strategy is found in the rational process models from Chapter 2, wherein Bayesian inference is approximated with sampling algorithms from machine learning and computer science. Resource constraints are specified within the Bayesian framework, and the resultant process models are still fundamentally Bayesian. Probability matching, for example, can be explained by assuming that sampling from the posterior distribution is costly (Vul et al., 2014). So for rational process models, the derived algorithmic models inherit the key attribute of Bayesian inference from the initial computational model. While algorithmic models will typically mirror the character of the initial computational model, it is not the rule. Fernbach and Sloman (2009), for example, start with a normative Bayesian model of learning causal structure. Based on deviations

from normative predictions, they develop a non-Bayesian heuristic model of structure learning that is based on local computations.

The second strategy of bridging levels has been applied to the take-the-best heuristic. Take-the-best can be used to choose between two alternatives that are characterized by a number of binary cues. The validity of a cue is defined as the percentage of correct inferences that result from using that cue to decide between alternatives. Take-the-best first orders cues according to their validity. Two options are compared by searching through these ordered cues, and search is terminated once a discriminatory cue is found. The heuristic is *non-compensatory*, since information from later cues cannot override the decision made based on the discriminating cue.

Take-the-best has been shown to be optimal with respect to various environmental structures (Dieckmann & Rieskamp, 2007; Katsikopoulos & Martignon, 2006; Martignon & Hoffrage, 1999, 2002). In general, take-the-best is successful in environments characterized by *diminishing returns* and with *correlated information* (Lee & Zhang, 2012). Diminishing returns refers to a skewed distribution of cue validities, with higher validity cues being followed by cues with much lower validities. Correlated information refers to cues that carry redundant information. In general, non-compensatory strategies are rational when most information is contained in the high-validity cues, which is the exactly the finding for take-the-best.

Computational analysis shows precisely the types of environments in which take-the-best performs well. Unlike the first bridging strategy, the final product gives no separate heuristic and computational model. Instead, one assumes the heuristic model is adaptive and the only task is to precisely characterize the environment to which it is adapted to (Gigerenzer & Todd, 1999). So on the fast-and-frugal approach, a bridge is formed between levels, but not between models. The computational analysis completely depends on the heuristic model. Answers to "why" questions are found only through close study of "what". Thus, inheritance from the initial model is even stronger under the second strategy of bridging levels.

3.3 A novel strategy for theory integration

Now which of the two bridging strategies is most promising for models of causal learning? The first strategy, using rational process models, would be straightforward. One simply needs to choose a preferred Bayesian model and then choose a sampling algorithm to approximate this model. I do not use this strategy. One reason why is that rational process models appear implausible as

mechanistic models since they make nearly the same representational assumptions as standard Bayesian models. In addition, there is reason to be skeptical that rational process models will be effective estimators of causal strength, which will be elaborated in section 3.6 below.

What about the second strategy? Might the proper analysis focus on a strong heuristic model, and then work out the computational implications? With a bit of reflection, it is easy to see that such a strategy will founder. Recall that the computational problem can be characterized by two components: the inferential goal and a formal description of the environment. In the take-the-best example, the inferential goal is clear: maximize the number of correct choices (e.g. to choose which of two German cities has a larger population). Thus, an essential aspect of the computational analysis is already incorporated into the heuristic model. Crucially, this aspect is missing from the problem of elemental causal induction. As we've seen, the inferential goal is precisely the point of contention across competing computational accounts of causal induction. So without a known inferential target, it is impossible to precisely characterize the performance of a given heuristic model.

The preceding discussion suggests why there is a paucity of work that attempts to bridge levels of analysis for models of causal learning. Fortunately, a novel strategy can be employed to resolve this impasse. The strategy can be summarized with a few general steps:

1. Find (or create) computational model $C(\beta_1, \dots, \beta_n)$ and algorithmic model $A(b_1, \dots, b_n)$.
2. Specify correspondence $\beta_1, \dots, \beta_n \rightleftarrows b_1, \dots, b_n$ between variables of the computational model and free parameters of the algorithmic model.
3. Work out implications to determine if step 2 gives useful insights.
4. Revisit first three steps as necessary.

The advantage of this approach is that it draws on the strengths of both levels of analysis. Namely, it preserves the strong justification of the computational model while maintaining the psychological plausibility of the algorithmic model. The next section uses this strategy to construct a bridge between the computational model of causal power and the algorithmic model of weighted ΔP.

3.4 A bridge between causal power and weighted ΔP

To motivate the connection between the power PC model and weighted ΔP, recall that generative causal power corresponds to the noisy-OR parameterization:

$$w_T = w_1 + w_0 - w_1 \times w_0$$

where $w_0 = P(e^+|c^-)$ and $w_T = P(e^+|c^+)$ and w_1 is the causal power for some cause *C*. Rearranging terms gives:

$$w_1 = w_T - (1 - w_1) \times w_0$$

The expression can be thought of as a 1-parameter weighted ΔP model with weight $(1 - w_1)$. On first inspection, this does not appear useful: to find w_1 requires already knowing w_1. But what if instead of knowing causal power the learner just has a reasonable guess for w_1?

There is considerable evidence that prior belief influences how people learn from contingency information (McKenzie & Mikkelsen, 2007). This prior belief forms the basis for the connection between causal power and the weighted ΔP model. Consider the 1-parameter weighted ΔP model:

$$w\Delta P = w_T - k \times w_0$$

where $0 \leq k \leq 1$. Now suppose that θ represents the prior expectation for causal power. Define $k = 1 - \theta$. This gives the model:

$$w\Delta P = w_T - (1 - \theta) \times w_0 \qquad (3.1)$$

Finally, the population quantities w_T and w_0 are typically not available, so instead sample estimates $\widehat{w}_T = \hat{P}(e^+|c^+)$ and $\widehat{w}_0 = \hat{P}(e^+|c^-)$ must be used. Below the estimates are also written as $\widehat{w}_T = w_T + \epsilon_T$ and $\widehat{w}_0 = w_0 + \epsilon_0$ with the errors both having an expectation of 0 (see Appendix B or C for a description of the error distribution). The weighted ΔP estimator is then:

$$w\Delta P = \widehat{w}_T - (1 - \theta) \times \widehat{w}_0$$

Weighted ΔP can now be assessed as an estimator of causal power. To begin, consider a generative context in which the base rate $\widehat{w}_0$ is equal to 1. In this instance, Cheng's (1997) power PC theory

asserts that there is no evidence to determine causal strength, and so people should withhold judgment. Accordingly, the causal power MLE is undefined at $\widehat{w}_0 = 1$. In contrast, weighted ΔP returns a judgment of θ, the prior expectation. This is exactly what the prior expectation is supposed to represent—belief before observing the evidence. It also matches with the Bayesian concept of a prior expectation.

Weighted ΔP can also be characterized with respect to its bias and variance. Let $w\Delta P = \widehat{w}_1$. Then the conditional bias is given by:

$$\begin{aligned} \widehat{w}_1 &= \widehat{w}_T - (1-\theta)\widehat{w}_0 \\ E[\widehat{w}_1|w_1] &= E[w_1 + w_0 - w_1 w_0 + \epsilon_T - (1-\theta)(w_0 + \epsilon_0)] \\ &= w_1 - w_1 w_0 + \theta w_0 + E[\epsilon_T] + (1-\theta)E[\epsilon_0] \\ &= w_1 + (\theta - w_1)w_0 \end{aligned}$$

So conditional on a causal power w_1, the weighted ΔP estimator is biased. Naturally, the bias is a function of the distance between the true value w_1 and the prior expectation θ. Bias is also an increasing function of the population base rate w_0. A population base rate of zero implies $\widehat{w}_0 = w_0 = 0$ and an unbiased weighted ΔP with $\widehat{w}_1 = \widehat{w}_T$.

Similar to a Bayesian estimator, weighted ΔP provides estimates that are often regressive to the prior expectation. For a Bayesian model, the amount of regression depends on the base rate of the effect and the sample size. With weighted ΔP, regression to the prior depends only on the base rate. To see how, recall that the causal power MLE is given by

$$\widehat{w}_{1.\mathrm{MLE}} = \frac{\widehat{w}_T - \widehat{w}_0}{1 - \widehat{w}_0}$$

Expanding the weighted ΔP estimator gives:

$$
\begin{aligned}
w\Delta P &= \widehat{w}_T - (1-\theta)\widehat{w}_0 \\
&= \widehat{w}_T - \widehat{w}_0 + \theta\widehat{w}_0 \\
&= \frac{\widehat{w}_T - \widehat{w}_0}{1-\widehat{w}_0}(1-\widehat{w}_0) + \theta\widehat{w}_0 \\
&= \widehat{w}_{1.\mathrm{MLE}} \times (1-\widehat{w}_0) + \theta \times \widehat{w}_0
\end{aligned}
$$

And so the $w\Delta P$ estimator is a weighted combination of the MLE and the prior expectation. As the sample base rate $\widehat{w}_0$ increases, more weight is given to the prior expectation. This seems reasonable since equation (2.4) shows that there is a quadratic increase in the MLE's variance with increasing w_0. A Bayesian estimator would weigh the combination by both the base rate and the sample size, which is sensible since the sample estimate is more reliable with larger *N*. Though it is rational to consider sample size, there is evidence that people in fact ignore it. For instance, Anderson & Sheu (1995) found contingency judgments to be insensitive to sample size. Not coincidentally, they were also early proponents of the weighted ΔP rule.

To facilitate comparisons with results presented in Chapter 2, again assume that the base rate w_0 is known and fixed. Then the mean-squared error of weighted ΔP is:

$$
\mathrm{MSE}[w\Delta P] = \big((\theta - w_1)w_0\big)^2 + \frac{w_T(1-w_T)}{N} \tag{3.2}
$$

This is the familiar $\mathrm{Bias}^2 + \mathrm{Variance}$ formula. If one has little prior information, a good choice of prior expected power in (3.2) is $\theta = \frac{1}{2}$ since it produces a maximum absolute bias of $|\theta - w_1| = 1/2$. It can be shown that the MSE of (3.2) only increases linearly as $w_0 \to 1$. It also has a finite maximum at $w_0 = 1$. This is in contrast to the variance for the MLE, which increases quadratically and has no upper bound as $w_0 \to 1$. Appendix C gives the derivation for (3.2). It also derives the MSE for a random base rate $\widehat{w}_0$, which again shows linear growth as the base rate increases.

Figure 3.1 compares the approximate MSE of the MLE versus weighted ΔP across various levels of causal power w_1 and at a sample size of $N = 10$. It shows what equations (2.5) and (3.2) tell us. The MLE and weighted ΔP are identical for $w_0 = 0$. As the base-rate w_0 grows, there is a fast increase in the MLE's mean-squared error while there is only a moderate increase for weighted ΔP. Also note that weighted ΔP performs better when bias is low in the $w_1 = 0.4$ and the

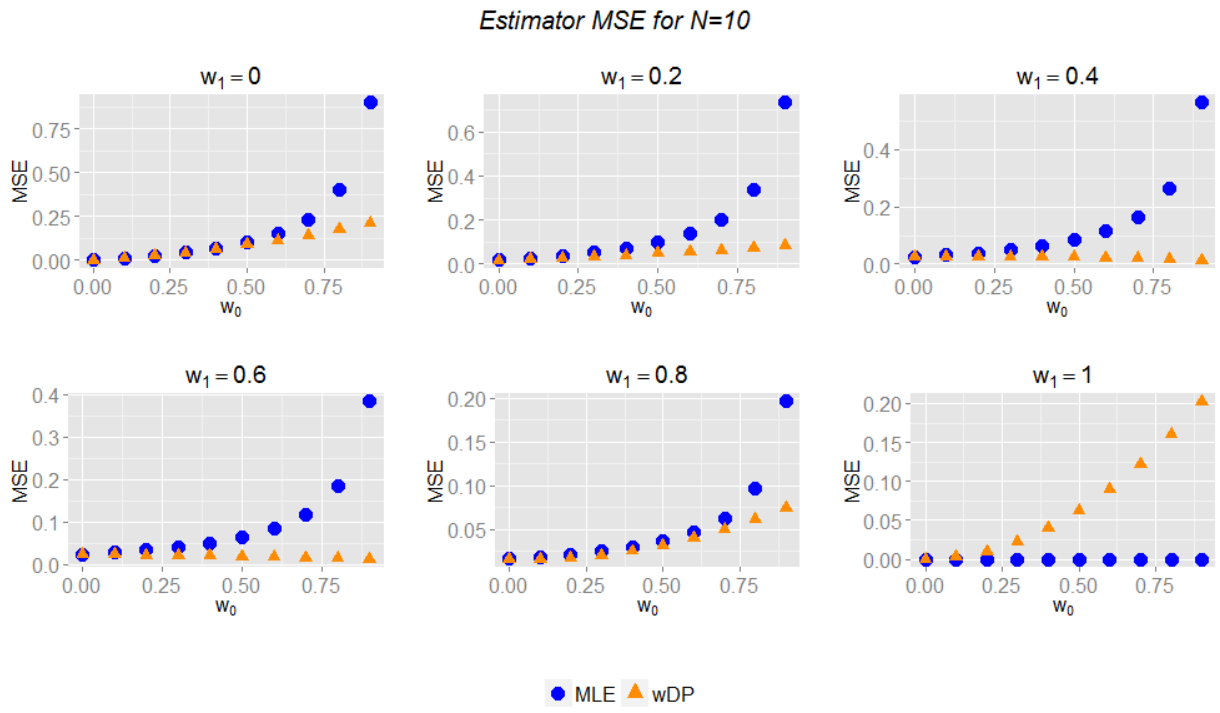


Figure 3.1. Mean-squared error of the causal power MLE (blue circles) and weighted ΔP (orange triangles) for a sample size $N = 10$, and with each panel showing a different level of causal power w_1. The weighted ΔP prior expectation is $\theta = 0.5$.

$w_1 = 0.6$ panels. The causal power MLE is superior to weighted ΔP only in the final panel with causal power $w_1 = 1$. In this case, the MLE always makes the correct prediction of $\widehat{w}_{1.\mathrm{MLE}} = 1$, while weighted ΔP will include error for positive base rates $w_0 > 0$.

Naturally, the relative performance of the maximum likelihood estimator improves with a larger sample size. Figure 3.2 compares estimators for a sample size of $N = 20$. As N increases, variance is reduced and the unbiased MLE converges to the true w_1. Variance is also reduced in the weighted ΔP estimator, but the bias is unaffected. The implication is that the MLE will have better relative performance with a larger N. In summary, weighted ΔP will be a biased estimator of causal power, though for small N or for large w_0, it will have a lower MSE than the causal power MLE.

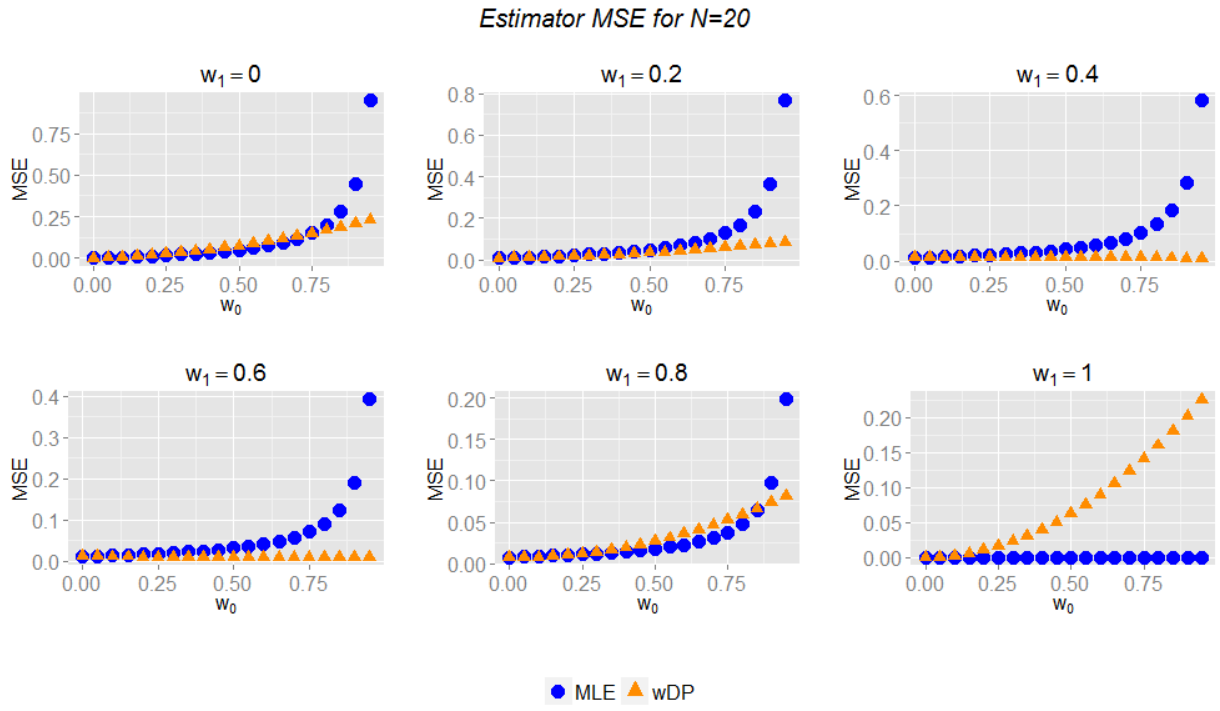


Figure 3.2. Mean-squared error of the causal power MLE (blue circles) and weighted ΔP (orange triangles) for a sample size $N = 20$, and with each panel showing a different level of causal power w_1. The weighted ΔP prior expectation is $\theta = 0.5$.

3.5 Conditional versus unconditional estimators

The focus thus far has been on causal judgment rules as they pertain to a single candidate cause C and its associated causal power w_1. In other words, the evaluation has been *conditional* on a given parameter w_1. Yet over our life history there are presumably many opportunities to make judgments over a diverse set of causes $\{C_1, \ldots, C_n\}$, each associated with distinct causal powers. As several authors observe, a strategy that is suboptimal on single occasions may be optimal on average (Brighton & Gigerenzer, 2008; Frank, 2013). Thus, it is also important to evaluate average performance over the long-run. This pertains to the *unconditional* properties of an estimator.

To evaluate performance over repeated samples, we must first characterize the population of causal powers. The population need not encompass all conceivable causes, but may rather refer to some more circumscribed set that is relevant to a given context, e.g., the set of causal powers that might be encountered in the scenario described in the experiment.

To begin, assume that candidate causal powers are identically distributed with population mean $E[w_1] = \theta$ and population variance $V[w_1] = \tau^2$. It was just shown that weighted ΔP is a conditionally biased estimator of causal power. What about its unconditional bias? The answer to this question is less clear cut. If a judge is well-calibrated to the environment, such that their prior expectation equals the population mean, then weighted ΔP will be unconditionally unbiased:

$$\begin{aligned} E\big[E[\hat{w}_1|w_1]\big] &= E[w_1 + (\theta - w_1)w_0] \\ &= \theta + w_0(\theta - E[w_1]) \\ &= \theta \end{aligned}$$

So weighted ΔP is unbiased when performance is averaged over repeated sampling of causes, i.e., sampling different values of w_1 from the population of causal powers. Even if the judge is not perfectly calibrated, unconditionally bias will be limited so long as they are close in their prior expectation to the population value.

Judgment rules can also be assessed by their unconditional mean-squared error. Derivations of the unconditional mean-squared errors for the MLE and weighted ΔP are found in Appendices B and C respectively. Figure 3.3 uses these derivations to compare estimator performance. The general message of Figure 3.3 mirrors what was shown for the conditional MSE. With increasing base rate the mean-squared error of the MLE grows rapidly while weighted ΔP has only gradual growth.

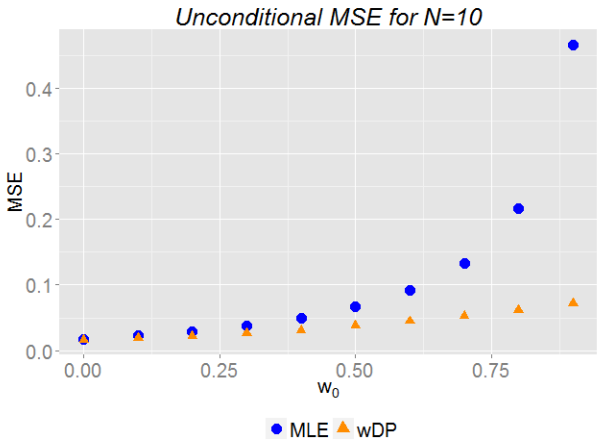


Figure 3.3. Plot of unconditional mean-squared error for the causal power MLE (blue circles) and weighted ΔP (orange triangles) for a sample size $N = 10$. Causal power is assumed to have a uniform random distribution, so the weighted ΔP rule with prior expectation θ=1/2 is unconditionally unbiased.

3.6 Weighted ΔP versus Bayesian Power, round 1

We have just seen that the weighted ΔP rule reduces error in high base rate contexts, but it is not the only game in town. In fact, we know when a better strategy exists. When model assumptions are met, the Bayesian posterior expectation will give the minimum mean-squared error (MMSE). Thus, the Bayesian model serves as a good benchmark for assessing model performance.

Cheng's original power PC theory was intended to describe only ordinal judgments of causal strength. The relative ordinal performance of various estimators will generally track with their mean-squared errors. However, it is illustrative to compare strategies by their success in ranking the strengths of candidate causes. This is done in the simulation study described below.

In the simulation presented below, cause C_1 is evaluated with respect to context B_1 by drawing a causal strength w_{11} and a background strength w_{01} from a joint uniform distribution. These quantities are combined using the Noisy-OR parameterization to find w_{T1}. Random draws are taken from the binomial distributions $B(N, w_{01})$ and $B(N, w_{T1})$ for the c_1^- and c_1^+ trials respectively. The resultant 2×2 contingency table is then used to form $\widehat{w}_{11}$ estimates. The procedure is repeated for cause C_2 and context B_2 to find $\widehat{w}_{12}$ estimates. Finally, the simulation experiment is performed over several different sample sizes *N*.

The study compares predictions of the causal power MLE, weighted ΔP with $\theta = 0.5$, and the Bayesian model with a joint uniform prior. Models are evaluated with respect to how often they predict the correct ordering of causal powers (e.g. how often $\widehat{w}_{11} < \widehat{w}_{12}$ when $w_{11} < w_{12}$). Note that for this simulation the Bayesian model will give the MMSE while weighted ΔP will be unconditionally unbiased.

Vul et al. (2014) argue that good decisions can be made from using just one or a few samples from the posterior distribution, especially when making a choice between only two alternatives. Consequently, a simple rational process model is included to test this claim. The rational process model draws 10 samples each from the posterior distributions of w_{11} and w_{12} using a Metropolis algorithm. The sample averages are then used to form its $\widehat{w}_{11}$ and $\widehat{w}_{12}$ estimates.

Results of the simulation study are presented in Table 3.1. The Bayesian model outperforms all other models, which is expected since it is normative for this problem. However, for small sample sizes the weighted ΔP and the Bayesian models exhibit very similar performance. The relationship between models changes as the sample size increases. At $N = 50$ and larger, the MLE

Table 3.1. Model accuracy in ordering of two causal powers. 10,000 simulations used for each sample size *N*.

	$N = 4$	$N = 8$	$N = 12$	$N = 16$	$N = 20$	$N = 50$	$N = 100$
MLE	.63	.68	.71	.74	.75	.83	.87
Weighted ΔP	.69	.72	.75	.76	.78	.82	.84
Bayes	.69	.73	.75	.78	.79	.84	.88
Rational process	.57	.60	.62	.63	.64	.68	.70

Notes. The rational process model uses the expectation of ten samples from the posterior probability distributions to order the two causes.

and the Bayesian model converge in performance while weighted ΔP begins to lag. This fits with the general fact that the Bayesian model will approach the MLE as sample size becomes large. The rational process model is the clear loser across all sample sizes. Performance of the rational process model could probably be improved with a better constructed sampling method, though it will still almost certainly be worse than weighted ΔP.

Next, Table 3.2 examines how often each model agrees with the Bayesian prediction across the various sample sizes. Weighted ΔP and the Bayesian model demonstrate high agreement across all sample sizes. In contrast, agreement between the MLE and the Bayesian model is modest for low *N*, and then gradually increases. Interestingly, the MLE and the Bayesian model exhibit higher agreement only for $N = 100$. The rational process model again fares considerable worse than the other models.

Table 3.2. Prediction agreement with Bayesian model. 10,000 simulations used for each sample size *N*.

	$N = 4$	$N = 8$	$N = 12$	$N = 16$	$N = 20$	$N = 50$	$N = 100$
MLE	.75	.80	.83	.85	.87	.91	.95
Weighted ΔP	.89	.92	.93	.94	.94	.92	.91
Rational process	.63	.65	.67	.68	.68	.71	.72

Notes. The rational process model uses the expectation of ten samples from the posterior probability distributions to order the two causes.

In summary, the Bayesian model exhibits a modest advantage over weighted ΔP when they are both used to rank the relative strength of two causes. This is despite the fact that the Bayesian model is considerably more complex. Overall, the simulation experiment reinforces the above findings that the weighted ΔP rule is an effective estimator of causal power.

3.7 Preventive causes

In this section I derive the weighted ΔP estimator for preventive causes. The derivation will show another important benefit of making an explicit connection between algorithmic and computational models. Specifically, I will use it to argue that previous formulations of the preventive model have been misconstrued.

The connection to causal power is motivated in a similar manner as above. First, begin with the Noisy-AND-NOT parameterization, which corresponds to preventive causal power:

$$w_T = w_0(1 - w_1)$$

Again, $w_0 = P(e^+|c^-)$ and $w_T = P(e^+|c^+)$ while w_1 is preventive causal power. Recall that under the Noisy-OR, w_1 could be expressed as a linear combination of w_T and w_0. With the Noisy-AND-NOT, no simple manipulation gives such a linear combination. Consequently, an indirect route must be taken to achieve this form.

By the rules of probability $P(e^+|c^-) = 1 - P(e^-|c^-)$ and $P(e^+|c^+) = 1 - P(e^-|c^+)$. Denote $\overline{w}_0 = P(e^-|c^-)$ and $\overline{w}_T = P(e^-|c^+)$. Then from the Noisy-AND-NOT:

$$\begin{aligned} w_T &= (1 - \overline{w}_0)(1 - w_1) \\ w_1 &= 1 - w_T - \overline{w}_0 + w_1\overline{w}_0 \\ w_1 &= \overline{w}_T - (1 - w_1)\overline{w}_0 \end{aligned} \tag{3.3}$$

And so w_1 is expressed as a linear combination, but of $P(e^-|c^+)$ and $P(e^-|c^-)$ instead of $P(e^+|c^+)$ and $P(e^+|c^-)$.

Before proceeding, it is helpful to reflect on the meaning of generative versus preventive causal power. Generative power describes causal influence when the candidate cause acts in isolation. In other words, it is just causal strength evaluated in a context in which the effect would never otherwise occur (but importantly, all enabling factors are present to allow the effect to

occur). Formally, generative power is the probability that the candidate cause produces the effect when no other generative causes are active, so in a context with $w_0 = 0$. The meaning of preventive power mirrors the generative definition, though it is a bit more convoluted. A cause's preventive power is the probability that it will prevent an effect that will otherwise occur. Hence, preventive power is evaluated with respect to a context in which the effect always occurs, or with $w_0 = 1$. Applying this definition within (3.3) gives $\overline{w}_0 = 1 - w_0 = 0$, which parallels the definition of generative power.

With an understanding of preventive power in place, we can proceed as before in forming a bridge between models. Once again, define the weight as $k = 1 - \theta$, where θ represents a prior expectation for preventive causal power. Substituting into (3.3) gives:

$$\begin{aligned} w\Delta P &= \overline{w}_T - k\overline{w}_0 \\ &= \overline{w}_T - (1-\theta)\times\overline{w}_0 \end{aligned} \tag{3.4}$$

So (3.4) gives the weighted ΔP model as an estimator of preventive causal power. The preventive model is essentially the same as the generative version. The only difference is that the focal outcome has been "reverse coded" from e^+ to e^-.

Though (3.4) is a simple modification, previous researchers have failed to appreciate its relevance. Instead, the standard practice for preventive causes has been to maintain e^+ as the focal event. Specifically, they will typically use the two parameter weighted ΔP model:

$$\begin{aligned} w\Delta P &= k_1 P(e^+|c^+) - k_2 P(e^+|c^-) \\ &= k_1 w_T - k_2 w_0 \end{aligned} \tag{3.5}$$

Some authors, such as Perales and Shanks (2007), fit a single set of weights for data that includes both generative and preventive conditions. Other authors allow different weights for generative and preventive conditions (e.g. Buehner et al. (2003)). Yet it is easy to show that different weights cannot accommodate weighted ΔP as an estimator for generative and preventive power. To see why, begin with (3.4), but then change the focal event from e^- to e^+:

$$\begin{aligned} w\Delta P &= \overline{w}_T - (1-\theta)\times\overline{w}_0 \\ &= (1-w_T) - (1-\theta)\times(1-w_0) \\ &= -[w_T - (1-\theta)w_0] + \theta \end{aligned}$$

Clearly, no choice of weights in (3.5) will give the preventive estimator. An intercept must be added to give the general linear combination model:

$$\widehat{w}_1 = k_0 + k_1 w_T + k_2 w_0 \tag{3.6}$$

The preceding derivations imply that different sets of coefficients are necessary to allow (3.6) to properly estimate generative and preventive power. The generative estimator could be represented with $k_0 = 0$, $k_1 = 1$ and $k_2 = -(1-\theta)$ and the preventive by $k_0 = \theta$, $k_1 = -1$, and $k_2 = (1-\theta)$.

Now instead of the one-parameter weighted ΔP model of (3.1) and (3.4), we could just use the linear combination model in (3.6) to explain and predict causal judgments. If all three weights are allowed to vary, the model would accommodate both generative and preventive forms while still having two additional free parameters to describe judgments. While this additional flexibility would seem desirable, it is possible for models to be too flexible. This issue will be explored further in Section 3.8.

Another strategy would be to use (3.6) and constrain the weights. The generative form would have constraints $k_0 = 0$ and $k_1 = 1$ while preventive weights would be $k_0 = \theta$ and $k_1 = -1$. With these constraints, the generative and preventive expressions of (3.6) are mathematically equivalent to generative and preventive weighted ΔP from (3.1) and (3.4). Yet the models may not be psychologically equivalent. In particular, (3.4) implies that the focal event—and so the focus of the judge—shifts from effect present events (e^+ events) to effect absent events (e^- events). In contrast, (3.6) suggests that the focus remains on e^+, but that the weighting of the outcomes changes. This distinction is subtle, but potentially testable. For instance, measurements could be made on how much time people spend examining e^+ versus e^- trials in generative versus preventive conditions. I suspect that the weighted ΔP representation better describes the underlying psychological process. Regardless, the weighted ΔP model is a more economical representation of the judgment rule, which is another reason it is preferred. And I will soon show

how this representation facilitates connections between weighted ΔP and associative models of causal learning.

One issue to consider is whether belief about generative and preventive strengths is the same. Suppose there is a generative scenario concerning cause C_g with strength w_g and a preventive scenario about cause C_p with strength w_p. With the exception of labeling one cause as "generative" and the other as "preventive", imagine that the scenarios are otherwise identical. In particular, the objective evidence across scenarios supports an inference of equal causal strengths $\widehat{w}_g = \widehat{w}_p$. For example, the generative scenario could have observed contingencies $\widehat{w}_T = 0.75$ and $\widehat{w}_0 = 0.25$ and so the preventive scenario would have $\widehat{w}_T = 0.25$ and $\widehat{w}_0 = 0.75$. If prior belief in causal strength is symmetric across generative and preventive causes, then judgments will be identical for these two hypothetical scenarios. But it is possible that people are asymmetric in their beliefs. Then the mere act of labelling a cause as generative or preventive will produce different judgments. If so, it will be necessary to distinguish between prior expectations for generative versus preventive causal strengths in (3.1) and (3.4) above. For instance, θ_g can denote the former and θ_p the latter. This is an empirical issue that will be taken up in the next chapter.

Finally, preventive power is often reported as a negative quantity on the $[-1,0]$ interval. A power of -1 then means that a cause will always prevent the effect. The weighted ΔP estimator can be used to find a negative preventive power, though some care must be taken to properly represent the weight. To find the proper expression, again begin with Noisy-AND-NOT and work forward to:

$$w_T = w_0(1 - w_1)$$
$$1 - \bar{w}_T = (1 - \bar{w}_0)(1 - w_1)$$
$$\bar{w}_T = w_1 + \bar{w}_0 - w_1 \times \bar{w}_0$$

Now we can map preventive power to $[-1,0]$ simply by reversing the sign of w_1. This gives:

$$\bar{w}_T = -w_1 + \bar{w}_0 + w_1 \times \bar{w}_0$$

Solving for power w_1 yields:

$$w_1 = -[\overline{w}_T - (1 + w_1)\overline{w}_0]$$

And substituting θ to represent prior expectation gives:

$$w_1 = -[\overline{w}_T - (1 + \theta)\overline{w}_0]$$

Naturally, the prior expectation θ is also on the $[-1,0]$ scale. Through some additional algebra, negative preventive power can also be expressed in terms of (3.6), the three parameter linear combination model:

$$w_1 = \theta + w_T - (1 + \theta)w_0 \tag{3.7}$$

So $k_0 = \theta$, $k_1 = 1$ and $k_2 = -(1 + \theta)$.

3.8 Model competition study

The preceding section gave an a priori argument for the proper functional form of preventive weighted ΔP. A question that naturally follows is whether the proposed changes make an empirical difference. This section addresses the empirical question with a replication of Perales and Shanks (2007) model competition study.

For their study, Perales and Shanks (2007) selected 114 conditions from 19 different published causal learning experiments. They included experiments that met the following criteria: 1) trial-by-trial presentation of covariation information 2) use of the standard causal probe question wording and 3) evaluation of only one candidate cause and one effect. Perales and Shanks argue that these criteria permit comparison among the largest number of models. Trial-by-trial presentation, for instance, allows for the inclusion of both associative and non-associative learning models.

Perales and Shanks (2007) used a cross-validation method to compare relative model performance. Cross-validation is a standard approach for comparing a diverse sets of models (see Hastie, Tibshirani, & Friedman (2009) for an in-depth treatment). Central to the approach is the distinction between true or population variation versus accidental or error variation. True variation is generated by the underlying causal mechanism of interest. Accidental variation is idiosyncratic

to a particular sample. Accordingly, true variation will generalize to new samples while accidental variation will not.

Models have varying degrees of flexibility. A highly flexible model can capture a wide variety of patterns in a data set. So if the true pattern is exotic, a flexible model will be able to describe such a pattern. A constant challenge is to use limited data to distinguish between true and accidental patterns. If a model is too flexible it will also fit accidental variation. This is sometimes referred to as *overfitting* the data.

A good model is one that is flexible enough to fit true variation, but not so flexible that it fits accidental variation. The challenge of finding such a model can be explained in terms of the *bias-variance trade-off* (Hastie et al., 2009). Earlier, we saw that an estimator's mean-squared error can be decomposed into $\text{Bias}^2 + \text{Variance}$. Flexible models are low bias since they can fit arbitrary patterns. But this also implies that they are high variance since flexible models will closely track any sample. In contrast, less flexible models will have lower variance, but they will often be biased. So to maximize predictive power, a model needs to strike a balance between bias and variance. The difficulty is that models are usually fit to a single sample of data while the question of prediction is left unanswered. Fortunately, cross validation provides a suitable method to address this difficulty.

The basic idea of cross-validation is to use a single sample of data to mimic a prediction task. The procedure is straightforward. First, the sample is randomly divided into a training sample and a test sample. Models are fit to the training sample. The fitted models are then used to make predictions on the test sample. This can be used to estimate the *test error* of each model, which is the quantity of interest. The procedure is repeated many times and the average from these iterations is used for the final test error estimate.

Perales and Shanks (2007) employed a 50-50 split in their cross-validation procedure, so half the data was used as a training sample and half as a test sample on each iteration. Of the many models they compared, they found that the EI rule had the lowest average test error, as measured by average mean-squared error. The weighted ΔP rule, for which they estimated a unitary set of $\{k_1, k_2\}$ weights across generative and preventive experiments, had a substantially worse fit.

Perales and Shanks' cross-validation study can be replicated to assess the weighted ΔP model described by equations (3.1) and (3.4) above. For the purposes of comparison, I will refer to (3.1) and (3.4) as *focal* $w\Delta P$ and Perales and Shanks' weighted ΔP model as *unitary* $w\Delta P$. The Bayesian

causal power model with a joint uniform prior is also added to the competition due to its previously observed strong empirical performance.

A summary of the findings is presented in Figure 3.4. The results replicate Perales and Shanks' findings for the EI rule (*aMSE* = 306.57, $w_a = .79, w_b = .54, w_c = .36, w_d = .31$) and for unitary $w\Delta P$ (*aMSE* = 434.42, $k_1 = 1.00, k_2 = .81$) in terms of average mean squared error and average estimated free parameters. The focal $w\Delta P$ rule (*aMSE* = 205.92, $\theta = .39$) had a substantially better fit than the EI rule, as well as all other models. Nonetheless, the Bayesian causal power model with a joint uniform prior also demonstrated very good fit (*aMSE* = 245.08). The similar performance of focal $w\Delta P$ and the Bayesian model is not surprising given that the models largely agree in their predictions.

The uniform prior Bayesian model provides good fit despite having no free parameters. However, this example illustrates why critics are skeptical of the lauding of Bayesian models as parameter free. Recall that preceding chapters examined a number of Bayesian models that varied with respect to their priors (uniform versus sparse and strong) as well as their generating functions (linear versus noisy-OR/noisy-AND-NOT). The SS prior had difficulty in describing key patterns of human judgments while the uniform prior model performed quite well. It was on this basis that the uniform prior model was included in the model competition while the SS prior model was omitted. Since previous data informed the selection of the Bayesian model, it is false to claim that it was constructed purely from a priori considerations. Though there are no free parameters,

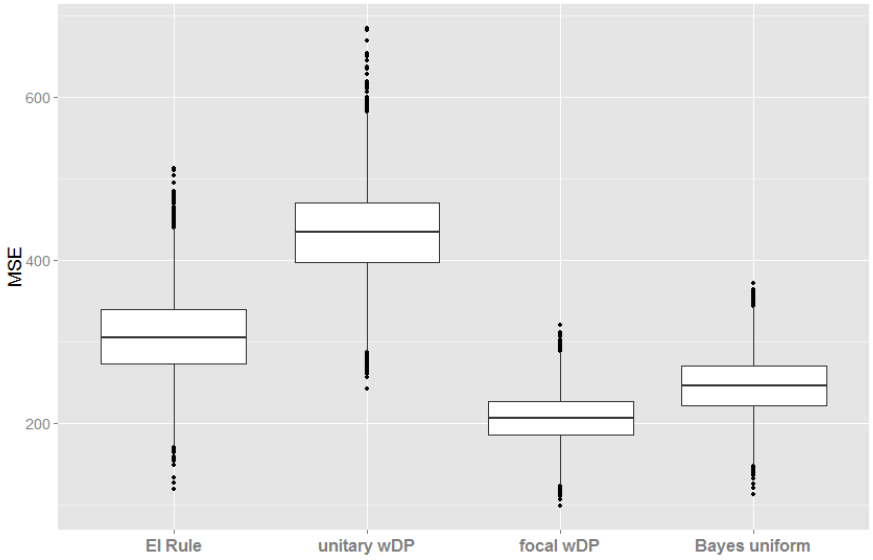


Figure 3.4. Boxplots of estimated test error over 10,000 cross-validation simulations.

considerable model fitting is still achieved through the choice of prior and generating function.

An additional point of interest concerns the reliability of performance for the different models. The boxplots in Figure 3.4 show a wide spread for the unitary $w\Delta P$ and, to a lesser extent, for the EI rule. This means that there are some test samples for which the models do relatively well while they are quite poor on others. Accordingly, $\mathrm{Var(MSE)} = 2604$ for the EI rule and $\mathrm{Var(MSE)} = 2974$ for unitary $w\Delta P$. Variance in MSE could result from model bias being more or less present on the randomly constructed training and test samples. The fluctuating influence of bias probably accounts for the highly variable performance of the unitary $w\Delta P$ model. If the above claims are correct, unitary $w\Delta P$ is suitable for generative causes, but misspecified for preventive causes. The upshot is that overall performance will be impaired by using generative and preventive conditions to estimate a single set of weights. Predictions will be especially poor for training samples with a high proportion of generative conditions and test samples with a high proportion of preventive conditions, and vice-versa.

Somewhat surprisingly, the variance in MSE for focal $w\Delta P$ ($\mathrm{Var(MSE)} = 880$) is lower than the variance for the Bayesian model ($\mathrm{Var(MSE)} = 1280$). This is despite the fact that focal $w\Delta P$ has a free parameter θ while the Bayesian model does not. This suggests that the free θ parameter is quite stable over training samples. In addition, there are probably certain subsets of conditions for which the Bayesian rule is more strongly biased. This possibility is explored further in Chapter 4.

Next, the cross-validation procedure is used to compare different weighted ΔP models. Two models are added to the analysis. The *dual* $w\Delta P$ model estimates two sets of weights, $\{k_{g1}, k_{g2}\}$ for generative conditions and $\{k_{p1}, k_{p2}\}$ for preventive conditions. This follows the approach of Buehner et al. (2003). Also included is the linear combination model of (3.6) with weights $\{k_{g0}, k_{g1}, k_{g2}\}$ for generative conditions and $\{k_{p0}, k_{p1}, k_{p2}\}$ for preventive conditions. As discussed above, the linear combination model is a more general form of focal $w\Delta P$. Therefore, it has the potential to overcome some of the bias of focal $w\Delta P$.

Figure 3.5 shows the cross-validation results. The dual $w\Delta P$ model ($aMSE = 333.42, k_{g1} = .98, k_{g2} = .60, k_{p1} = .98, k_{p2} = .60$) has better average test error than unitary $w\Delta P$ ($aMSE = 434.60, k_1 = 1.00, k_2 = .81$). However, the dual model is still much worse than focal $w\Delta P$ ($aMSE = 205.78, \theta = .39$). Crucially, simply allowing a different set of weights across

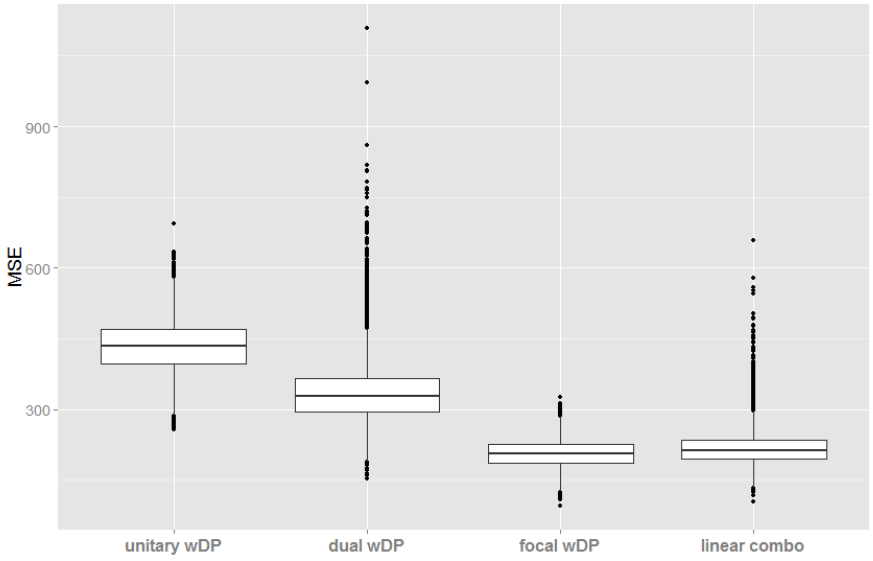


Figure 3.5. Boxplots of estimated test error over 10,000 cross-validation simulations.

generative and preventive conditions is not sufficient for good model fit. These findings further support the claim that preventive weighted ΔP models have previously been misspecified.

The linear combination model ($aMSE = 217.14,\ k_{g0} = -.09, k_{g1} = 1.10, k_{g2} = -.58,$ $k_{p0} = -.51, k_{p1} = .96, k_{p2} = -.36$) is also worse than focal $w\Delta P$. This is probably because the linear combination model has done little to reduce bias while the added flexibility has increased variance. Since variance is necessarily right-skewed, the result is to increase the average mean-squared error. Importantly, the average estimated parameters of the linear combination model are close to the constraints implied by focal $w\Delta P$ with generative weights $k_{g0} \approx 0$, $k_{g1} \approx 1$ and preventative weights $k_{p1} \approx 1$ and $k_{p2} \approx -(1 + k_{p0})$, since preventive power was recorded as negative. If focal $w\Delta P$ is the true data-generating model, then the linear combination estimates should approach the implied constraints as the number of included conditions and observations are increased. Convergence to the constraints should also be faster if the included conditions constitute a balanced representation of the condition space.

In summary, forming an explicit connection between computational and algorithmic accounts leads to a novel insight about the functional form for preventive weighted ΔP. Though the derivation and expression are simple, this detail has eluded researchers for some time. Indeed, in testing their "Cell A heuristic", Arkes and Harkness (1983) express great surprise that strength ratings did not increase monotonically with the frequency of what we've referred to as (e^+, c^+)

trials. Instead, strength judgments form a check-mark shaped pattern when moving from negative to positive contingency. This is exactly the pattern predicted by the focal weighted ΔP model. In other words, focal $w\Delta P$ implies that a "Cell A" heuristic (e^+ as focal) should shift to a "Cell B" heuristic (e^- as focal) when the context is preventive with a negative contingency.

3.9 Unknown causal direction

For completeness, I discuss a weighted ΔP estimator when causal direction is unknown. With unknown direction, one learning strategy is to simultaneously look for evidence of a generative and a preventive cause, and then choose the direction with the larger magnitude. Such a strategy seems unlikely in terms of stimulus processing. In the associative models described below, it would require that each trial be double coded.

Much preferable is a weighted ΔP rule for which generative and preventive forms agree so that the choice of e^+ or e^- as focal does not matter. Does such a strategy exist? In fact, with $\theta = 0$ the generative and preventive forms do coincide:

$$\begin{aligned} w\Delta P &= w_T - (1-\theta)w_0 \\ &= w_T - w_0 \\ &= -[(1-w_T) - (1-w_0)] \\ &= -[\overline{w}_T - \overline{w}_0] \\ &= -[\overline{w}_T - (1-\theta)\overline{w}_0] \end{aligned}$$

So both forms simply give the standard ΔP rule.

The choice of $\theta = 0$ appears intuitively reasonable. If one is uncertain of causal direction, a seemingly sensible characterization of this is with the prior belief that the cause is just as likely to either increase or decrease the probability of the effect. Hence, $P(e^+|c^+) > P(e^+|c^-)$ seems just as plausible as $P(e^+|c^+) < P(e^+|c^-)$. Or in weights notation, $w_T > w_0$ just as plausible as $w_T < w_0$. If these beliefs are symmetric and they "cancel out", then prior belief should give $w_T \approx w_0$. This is reflected in the model since $w_T = w_0$ is consistent with a prior belief of $\theta = 0$. To see why, observe that:

$$w\Delta P = w_T - (1-\theta)w_0$$
$$= w_T - (1-\theta)w_T$$

when $w_T = w_0$. So in this case, the updated estimate $w\Delta P$ will match the prior expectation only when $\theta = 0$.

It is an empirical question whether people in fact use the standard ΔP rule when causal direction is unknown. People may apply ΔP across all trials, or they may use ΔP on an early sample of trials to establish a direction, and then weighted ΔP is applied to later trials. This is an interesting question for future research.

3.10 Building more bridges: weighted ΔP as a Rescorla-Wagner process model.

When making causal judgments, it is questionable whether people directly represent the conditional probabilities $P(e^+|c^+)$ and $P(e^+|c^-)$. As has been discussed, the fact that judgments agree with some model does not imply that people explicitly follow that model or use the same representations. Indeed, Wasserman et al. (1993) found that a person's subjective probability ratings for $P(e^+|c^+)$ and $P(e^+|c^-)$ did not well predict their causal judgments. This led them to prefer Rescorla-Wagner as a process model. Shanks (1985, 1987, 1995) also favors associative over rule-based models as descriptions of psychological processes.

No choice of parameters for the conventional Rescorla-Wagner (R-W) model can be chosen so that the weighted ΔP of (3.1) and (3.4) are among its equilibria (Wasserman et al., 1993). Instead, it is necessary to modify R-W to a more general form. This section elaborates two ways that this may be achieved. First, recall from Chapter 1 that R-W can be expressed as:

$$\Delta V_i^t = \begin{cases} \alpha_i\beta^+\left[\lambda - \sum V_j\right] & \text{if both } C_i \text{ and } E \text{ appear in trial } t \\ \alpha_i\beta^-\left[0 - \sum V_j\right] & \text{if } C_i \text{ appears and } E \text{ does not appear in trial } t \end{cases}$$

Now consider an elemental causal induction problem with candidate cause C, background cause B and effect E. Denote association strengths for the candidate and background causes as V_C and

V_B, respectively. Suppose $\alpha_0 = \alpha_1$, $\beta_0 = \beta_1$ and a maximum possible association strength of $\lambda = 1$, which are all standard assumptions. One way to generalize R-W is with the introduction of an "attenuation" parameter κ, which is described in the next section.

3.10.1 κ-attenuation model

Introduce an additional parameter κ with $\kappa \in [0,1]$. The parameter can be given various interpretations. For starters, we can say that κ describes the attenuation of attention given to the background strength V_B on cause present trials. More simply, κ is a measure of base rate neglect. The κ parameter is incorporated into the R-W model with a neglect function $g(C)$, which is given by:

$$g(C) = \begin{cases} \kappa & \text{for } c^+ \text{ trials} \\ 1 & \text{for } c^- \text{ trials} \end{cases}$$

The neglect function multiplies the background weight V_B. With $\kappa = 0$ there is total neglect of background strength while with $\kappa = 1$ there is no neglect. Following Chapman & Robbins (1990) it is easy to show that the modified model will give the weighted ΔP rule at equilibrium (see Appendix D). The equilibrium association strengths are equal to:

$$V_C = P(e^+|c^+) - \kappa \times P(e^+|c^-)$$
$$V_B = P(e^+|c^-)$$

Suppose that neglect of the base rate is $\kappa = (1 - \theta)$ where θ represents a prior expectation for causal strength. Then at equilibrium the same desirable properties are achieved that were shown above for the weighted ΔP estimator. Also note that the equilibrium background strength is equal to the objective conditional probability with $V_B = P(e^+|c^-)$. We will see that this is a distinguishing prediction when compared to the modified R-W model proposed in the next section.

3.10.2 Unequal association strength model

Another way to obtain the weighted ΔP equilibrium is to allow the maximum association strengths (the λ's) to differ across contexts. Denote λ^+ as the maximum association strength for the c^+ context and λ^- as the maximum association strength for the c^- context. Furthermore, assume $\lambda^+ =$

1 and $\lambda^- = \lambda$ with $0 < \lambda < 1$. Then it can be shown (see Appendix D) that the equilibrium association strengths are equal to:

$$V_C = P(e^+|c^+) - \lambda \times P(e^+|c^-)$$
$$V_B = \lambda \times P(e^+|c^-)$$

Again if we set $\lambda = (1 - \theta)$, then weighted ΔP is the equilibrium with θ interpreted as the prior expected strength. However, now the predicted background strength does not equal the objective conditional probability. Instead, its equilibrium value is $\lambda \times P(e^+|c^-)$. Thus, the two modified R-W models have different predictions for the strength attributed to the background cause. This is one testable prediction that may be used to discriminate between the two models.

Another difference between models will be found in their speed of convergence to equilibrium. The augmented kappa model influences all c^+ trials while the lambda model only influences the (c^-, e^+) trials. One would expect, then, that the kappa model will converge to equilibrium more quickly. Furthermore, this discrepancy should be exaggerated for data with few (c^-, e^+) trials.

3.11 Dynamical weighted ΔP rule

With enough learning trials, Bayesian models will eventually converge to the causal power sample estimate while the weighted ΔP rule will remain biased. This bias is generally consistent with the behavioral evidence, where the typical finding is that judgments reach equilibrium after about 20 learning trials. An important detail is that causal learning experiments are administered on a single occasion. Yet most tasks require repeated exposure over multiple occasions before large learning gains accrue. Thus, it is possible that judgments may continue to evolve over time with repeated learning opportunities. One weakness of the weighted ΔP rule is that once $\widehat{w}_T$ and $\widehat{w}_0$ approach their population quantities, no further evolution of belief is possible.

Fortunately, a simple augmentation to weighted ΔP allows for more substantial changes in causal belief. Define the dynamical weighted ΔP rule as:

$$\widehat{w}_{1(0)} = \theta$$
$$\widehat{w}_{1(i+1)} = \widehat{w}_{T(i+1)} - \big(1 - \widehat{w}_{1(i)}\big)\widehat{w}_{0(i+1)}\,, \qquad i = 0, \ldots N \tag{3.8}$$

In the above expression, $\widehat{w}_{T(i)}$ and $\widehat{w}_{0(i)}$ are the respective estimates of $P(e^+|c^+, b^+)$ and $P(e^+|c^-, b^+)$ after the i^{th} occasion. The $\widehat{w}_{1(i)}$ is similarly defined for $P(e^+|c^+, b^-)$ while $\widehat{w}_{1(0)}$ is the prior expectation for w_1. What constitutes an "occasion" is an open question. It could be a single learning trial or an entire learning experiment.

An attractive property of dynamical weighted ΔP is that it will eventually converge to the true causal power w_1. To see why, first suppose that $\widehat{w}_0$ and $\widehat{w}_T$ have converged to their true values w_0 and w_T so that true power is given by $w_1 = \frac{w_T - w_0}{1 - w_0}$. Also suppose that $\widehat{w}_{1(i)} \neq w_1$ so that $\widehat{w}_{1(i)} = w_1 + \epsilon$. Now use (3.8) to find the updated $\widehat{w}_{1(i+1)}$:

$$\begin{aligned}\widehat{w}_{1(i+1)} &= w_T - \big(1 - \widehat{w}_{1(i)}\big)w_0 \\ &= w_T - \big(1 - (w_1 + \epsilon)\big)w_0 \\ &= w_T - w_0 + w_1 w_0 + \epsilon w_0\end{aligned}$$

The error for $\widehat{w}_{1(i+1)}$ is given by:

$$\begin{aligned}\widehat{w}_{1(i+1)} - w_1 &= (w_T - w_0 + w_1 w_0 + \epsilon w_0) - (w_T - w_0 + w_1 w_0) \\ &= \epsilon w_0\end{aligned}$$

So the error goes from $\widehat{w}_{1(i)} - w_1 = \epsilon$ to $\widehat{w}_{1(i+1)} - w_1 = \epsilon w_0$. Note that $0 \leq w_0 \leq 1$. If $w_0 = 0$, then convergence occurs in one step. If $w_0 = 1$ then $\epsilon_{(i+1)} = \epsilon_{(i)}$ and $\widehat{w}_1$ never converges. For $0 < w_0 < 1$ we have $\epsilon w_0 < \epsilon$ and $\widehat{w}_{1(i+1)}$ is closer to the true value than $\widehat{w}_{1(i)}$. With an additional update step the error will be:

$$\begin{aligned}\epsilon_{(i+2)} w_0 &= (\epsilon w_0) w_0 \\ &= \epsilon (w_0)^2\end{aligned}$$

By induction, after the n^{th} update the error will be $\epsilon(w_0)^n$. Hence, (3.8) converges to the true w_1 since $(w_0)^n \to 0$ as $n \to \infty$. So the dynamical weighted ΔP rule, like Bayesian causal power, also converges to the true value.

The frequency with which $\widehat{w}_1$ is updated can be thought of as a tuning parameter. More frequent updates will result in quicker convergence to the sample estimate, and so the estimator will be less biased. Yet this reduction in bias will come at the price of increased variance.

Other strategies may also be used for tuning the dynamical weighted ΔP model. For instance, instead of replacing $\widehat{w}_1$ with the new estimate on each occasion, a weighted average could be taken so that:

$$\widehat{w}_{1(\mathrm{i}+1)} = (1-\tau)\times\widehat{w}_{1(\mathrm{old})} + \tau\times\widehat{w}_{1(\mathrm{new})} \tag{3.9}$$

Where $\widehat{w}_{1(\mathrm{old})}$ and $\widehat{w}_{1(\mathrm{new})}$ are the previous and new estimates, respectively. And $\tau \in [0,1]$ is the constant tuning parameter. So as $\tau \to 1$ bias is reduced and variance increases. The right choice of τ could allow dynamical weighted ΔP to better approximate Bayesian models.

Figure 3.6 compares the standard weighted ΔP rule, dynamical weighted ΔP, and Bayesian causal power. Model predictions were made for 200 learning trials generated from a Noisy-OR parameterization with $w_0 = 0.2$ and $w_1 = 0.7$. For the dynamical rule, the tuning parameter of equation (3.9) is set to $\tau = 0.5$.

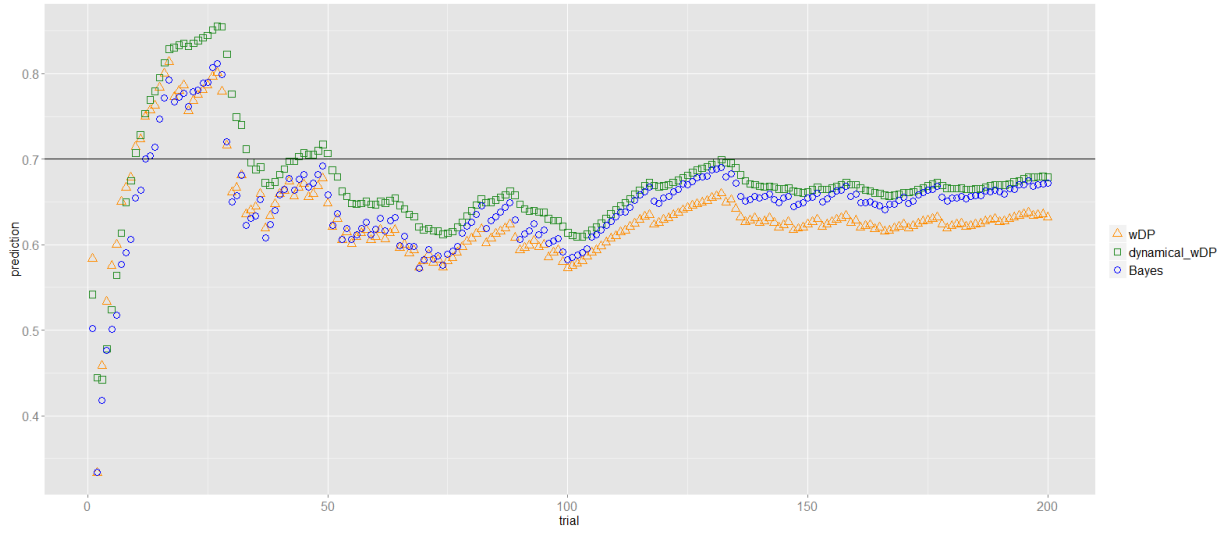


Figure 3.6. Model predictions of weighted ΔP (orange triangles), dynamical weighted ΔP with $\tau = 0.5$ (green squares) and Bayesian causal power (blue circles). Prior expected power is 0.5 for both weighted ΔP models. Learning data was randomly generated from a Noisy-OR parameterization with $w_0 = .2$ and $w_1 = .7$ (black line). The learning data contain an equal number of cause-present and cause-absent trials.

Naturally, each estimator is high in variance for the early trials, though dynamical weighted ΔP appears to be the highest in variance. Once evidence accumulates and the estimators settle down, and the Bayesian prediction is bracketed by the two weighted ΔP models for much of the range. At the beginning of the learning trials, Bayesian predictions are more similar to static weighted ΔP, while Bayes and the dynamical rule become more similar towards. For these data the dynamical rule would probably better approximate the Bayesian model by setting the tuning parameter to $\tau < 0.5$.

The dynamical weighted ΔP rule leads to a key insight regarding the process models discussed in the previous section. Namely, the same type of updating can be used in the process models so that Rescorla-Wagner converges to causal power (see Appendix D). Danks, Griffiths, and Tenenbaum (2003) also describe a R-W model that converges to power. This is achieved by incorporating the Noisy-OR or the Noisy-AND-NOT prediction into the model. It turns out that under certain assumptions the κ-attenuation model is identical to the Danks et al. (2003) model. In particular, if the κ parameter is updated after every trial, then the κ model is identical to the noisy-OR prediction model for generative causes. This is shown in Appendix D. There is also a near identity for between the κ model and the noisy-AND-NOT for preventive causes, thought this relationship is a bit more complex.

The connection between the weighted ΔP model and the Rescorla-Wagner model is important for a number of reasons. First, it provides a more specific account of the potential operations used to implement the weighted ΔP rule. In addition, the connection relates ideas concerning the bias-variance trade-off to process level theories of causal learning. Previously, when the data has been found to disagree with R-W predictions, one strategy has been to adjust the learning rates and then claim that observed judgments had not yet reached equilibrium. The current findings provide a competing explanation: judgments will not reach equilibrium because people are using a biased R-W model. Further, this biased rule can be understood as a rational strategy when the goal is to minimize the mean-squared error of prediction.

3.12 Model uncertainty

Until now, causal power has been assumed to be the true generative model. Cheng (1997) gives a strong a priori argument to justify causal power. That is, she demonstrates that causal power is the

correct measure of causal strength when model assumptions are met. But in practice, how often will the structure of the environment reflect the assumptions of causal power?

The utility of a given model will naturally depend on its environment of application. Environments may be information rich or poor. An information rich context is one in which the judge can be confident that the candidate cause is independent of all alternative causes. Rich contexts also allow for many observations. Good scientific experiments, with independence enforced through control and randomization, and additional observations available through replication, provide the gold standard for learning contexts. Yet everyday learning environments are often information poor in that they are purely observational with only a few sampled outcomes.

Another complication is that the inferential goal will also be context dependent. Whether it is better to extract association strength versus causal power will depend both on the context and on how the knowledge is to be applied. Research on Bayes nets emphasizes the distinction between predictions from observations versus predictions from interventions (Hagmayer, Sloman, Lagnado, & Waldmann, 2007). Return to the example of the association between white hair and heart disease from Chapter 1. The association will be useful to a paramedic needing to assess the condition of an unresponsive patient. In contrast, this knowledge is of no use to the cardiologist—intervening on white hair by dyeing or shaving it will not cure heart disease. The key question, then, is what role do people typically assume in their everyday contingency judgments? Are they more often paramedics or cardiologists? Or perhaps people change perspectives in accordance with the evidence and their inferential goals?

Causal knowledge is esteemed because of the special type of prediction it affords, namely predictions from interventions (Pearl, 2009). Cheng's causal power theory gives predictions for how a causal intervention will influence the outcome in a novel environment. However, it is often the case that the learner is consigned to an observational role only. In such contexts, the discovery of "mere association" may be the only reasonable inferential target.

Recall that a crucial assumption of causal power is that the background and candidate causes exert independent influence on the effect. Is there reason to believe that independence is typical of the causal systems that people encounter? It would seem that independence is rarely found in observational, naturalistic settings. Indeed, the fields of econometrics and structural equation modeling were created to extract causal information from observational data. Confounds regularly plague causal inference in the sciences. One example comes from the study of how maternal age

causally influences autism. Early research suggested that advanced maternal age did cause an increased risk for autism. However, many of these studies were limited in their control of important confounds (Croen, Najjar, Fireman, & Grether, 2007). For instance, older mothers tend to couple with older fathers, and a father's age has been found to be an important predictor of autism risk (Croen et al., 2007; Sandin et al., 2015). So in this example, assuming that maternal age is independent of all alternative causes will likely bias any estimate of its causal strength.

To be sure, Cheng (2000) suggest that independence is a useful assumption when reasoners do not have adequate information to assess potential interactions of alternative causes. Independence assumptions have proved valuable in other areas. For example, the Naïve Bayes classifier assumes that object features are independent conditional on class membership. Despite this simplifying assumption, Naïve Bayes has been shown to outperform more sophisticated alternatives (Hastie et al., 2009).

The question, then, is whether the independence assumption is a generally useful one. The difficulty is that for causal learning, and for cognitive science more generally, it is not clear what environments should be used to test this assumption (Jones & Love, 2011). Is it the ancestral environment in which our cognitive capacities evolved? Or is it the contemporary environments in which these reasoning abilities develop and are now employed? Absent knowledge about the relevant environment, the best that can be done is to explore how competing models perform over heterogeneous causal structures. A limitation of this approach is that it still requires assumptions about the distribution of these structures. So a potential future strategy is to characterize real causal environments and then evaluate how various strategies perform in them. This possibility is explored more closely in Chapter 5. But first, a simulation study is used to examine the influence of uncertain environments.

3.12.1 Relaxing the independence assumption

This section explores the implications of relaxing the independence assumption in the power PC model. Recall that generative causal power corresponds to the noisy-OR parameterization:

$$w_T = w_1 + w_0 - w_1 \times w_0$$

A more general parameterization is given by:

$$w_T = w_1 + w_0 - w_{1|0} \times w_0 \tag{3.10}$$

where the $w_{1|0}$ term does not refer to any edge weight, but instead is used to describe an interaction between the candidate causal strength w_1 and background strength w_0. A large range of causal models can be represented by (3.10) with $0 \leq w_{1|0} \leq 1$. At the extremes, $w_{1|0} = 0$ gives the ΔP rule while $w_{1|0} = 1$ gives the Per Cent Success rule. Of course, causal power is given when $w_{1|0} = w_1$. Strong positive interactions are not allowed as $w_{1|0} \in [0,1]$ implies that $w_T \leq w_1 + w_0$. Appendix A gives a general account for dependence between the candidate cause *C* and background *B*.

Begin by assuming that causal power holds on average but that there are deviations from independence. Formally, assume that $w_1 = w_{1|0} + \epsilon$ with $E[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2$. To isolate the influence of model uncertainty, assume that the causal strengths w_T and w_0 are known constants. Then the causal power MLE is unbiased and the MSE is equal to the variance with:

$$\mathrm{Var}[\widehat{w}_1] = \left(\frac{w_0}{1 - w_0}\right)^2 \sigma^2$$

So the MSE is a function of the base rate of the effect, with a rapid increase as $w_0 \to 1$. This result is similar to what was found in Chapter 2, with causal power the true model and with uncertainty for the causal strengths.

The weighted ΔP estimator naturally extends to (3.10) by writing:

$$w_1 = w_T - \left(1 - w_{1|0}\right) w_0$$

And so the weighted ΔP estimator is:

$$\widehat{w}_1 = w_T - (1 - \theta) w_0 \tag{3.11}$$

where θ is now a prior expectation for the interaction term $w_{1|0}$. The bias of (3.11) is given by:

$$E[(\widehat{w}_1 - w_1)] = w_0(\theta - w_1)$$

So bias is an increasing function of the base-rate of the effect. Without any background information, the safest prior expectation will be $\theta = 0.5$ as it will result in a maximum Bias^2 of $(0.5)^2 = 0.25$.

The MSE of weighted ΔP is:

$$E[(\hat{w}_1 - w_1)^2] = w_0^2[(\theta - w_1)^2 + \sigma^2]$$

And it is the familiar $\text{Bias}^2 + \text{Variance}$ formula.

The above results can be used to compare the causal power MLE and weighted ΔP in the context of model uncertainty. When $w_0 = 0$, the MSE of both estimators is 0 since there is no interaction between w_1 and w_0, and so no disturbance term. For $w_0 > 0$, weighted ΔP will have a lower MSE when:

$$\text{MSE}_{\text{w}\Delta\text{P}} < \text{MSE}_{\text{MLE}}$$

$$w_0^2[(\theta - w_1)^2 + \sigma^2] < \left(\frac{w_0}{1 - w_0}\right)^2 \sigma^2$$

$$(\theta - w_1)^2 < \left[\frac{1}{(1 - w_0)^2} - 1\right] \sigma^2$$

Not surprisingly, weighted ΔP is a good estimator relative to causal power when:

- Bias is low
- Average deviation from power is high (i.e. σ^2 is high)
- The base rate of the effect w_0 is high.

Hence, weighted ΔP may be a good alternative to causal power when the underlying generative model is unknown.

3.12.2 Simulation study

In the formal characterization thus far, Bayesian causal power and the causal power MLE are identical. This is because of the assumption that w_T and w_0 are known. This is equivalent to assuming a sample size that approaches infinity, which implies that the Bayesian estimate converges to causal power. What about the case when there is uncertainty in the w_T and w_0 as well

as an uncertain interaction $w_{1|0}$? In this situation, the Bayesian model will differ from the causal power MLE. However, since there is no closed form expression for the Bayesian estimate it cannot be explicitly compared to weighted ΔP. Instead it becomes necessary to use simulation results to evaluate model performance.

The simulation study below draws w_0 and w_1 from a random uniform distribution. An interaction term $w_{1|0}$ is drawn from a beta distribution with expectation $E[w_{1|0}] = w_1$. So once again, causal power is on average the correct model. The details of how $w_{1|0}$ is sampled are provided in Appendix E.

Next, w_T is found using equation (3.10). The w_0 and w_T parameters are then used to generate samples from a binomial distribution, producing a 2x2 contingency table. Weighted ΔP and Bayesian causal power are applied to the contingency table to form causal strength estimates. The Bayesian model uses a joint uniform prior over (w_0, w_1) while weighted ΔP uses a prior expectation of $\theta = 0.5$. The simulation was repeated over various sample sizes N.

Results are plotted in Figure 3.7 below. For small sample sizes, the Bayesian model strongly outperforms weighted ΔP. This is because weighted ΔP is a high variance estimator for small N as it uses the sample estimates $\widehat{w}_T$ and $\widehat{w}_0$. The Bayesian model, in contrast, smooths the sample estimates by combining them with prior information. Performance of both estimators improves with more sample information, though weighted ΔP improves more rapidly. So for moderate sized samples, the two estimators perform similarly.

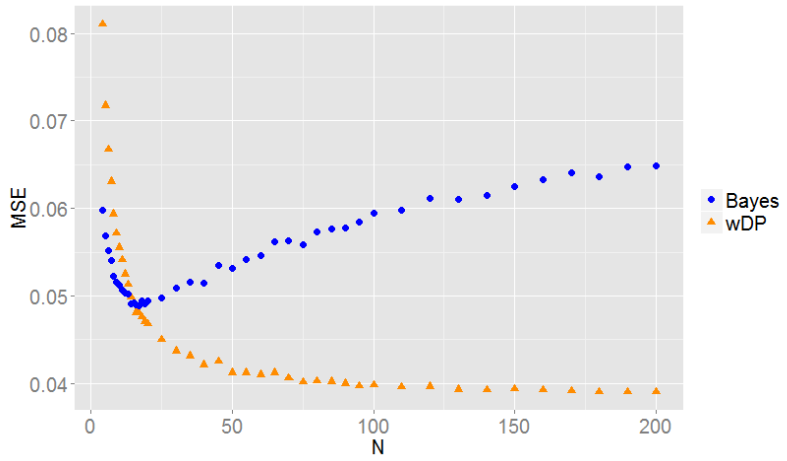


Figure 3.7. Mean-squared error against sample size for the Bayesian uniform and weighted ΔP models. The figure depicts relative performance in the context of parameter and model uncertainty. 10,000 simulations were performed at each sample size N.

As sample information accumulates, weighted ΔP continues to improve, though at a slower rate. Interestingly, the Bayesian model reaches a minimum and then performance degrades with larger samples. Why is this so? In fact, this is another example of the bias-variance trade-off. As the sample size increases the Bayesian model converges to the causal power MLE. We know from above that the MLE is unbiased on average, but this comes at a cost of increased variance. With model uncertainty we will sometimes have $w_{1|0} \gg w_1$ or $w_{1|0} \ll w_1$ and the MLE will badly miss the true value. In contrast, the weighted ΔP model is biased since it does not converge to power. Clearly, this property gives weighted ΔP an advantage over Bayesian power when *N* is large.

To be sure, an optimal Bayesian model could be constructed for this problem. But this would require an additional parameter for the interaction term $w_{1|0}$, and so belief would be represented over a 3-dimensional parameter space of $(w_0, w_1, w_{1|0})$. This would certainly constitute a marked increase in model complexity while improvement over the weighted ΔP rule would probably be negligible for moderate to large sample sizes.

3.13 Summary

To summarize the chapter, weighted ΔP has been shown to be an effective estimator of causal power when there is uncertainty in the observed probabilities $P(e^+|c^+)$ and $P(e^+|c^-)$. And more generally, it is a good estimator of causal strength when there is uncertainty about the interaction between the candidate and background causes. Weighted ΔP achieves performance comparable to Bayesian models while making far fewer representational commitments. Indeed, the simplicity of weighted ΔP allowed for its characterization as a Rescorla-Wagner type process model that requires only the incremental adjustment of association strengths.

The replication of Perales and Shanks (2007) meta-analysis found weighted ΔP to have better average fit to human judgments than their preferred EI Rule. Weighted ΔP was also better than the Bayesian uniform model, though they were much closer in performance. This is not surprising, as the models give similar predictions. Much work still needs to be done to assess the empirical performance of weighted ΔP, which is the primary task of Chapter 4.

Chapter 4.

Empirical investigation

4.1 Introduction

The weighted ΔP and uniform prior Bayesian models give similar predictions, which suggests that they will have similar empirical performance. Yet in the Chapter 3 replication of Perales and Shanks (2007), weighted ΔP consistently outperformed the Bayesian model. The primary goal of this chapter is to find important differences in model predictions. This will hopefully allow for a stronger empirical contrast between the leading model contenders.

4.2 Critical comparisons

Recall that causal learning conditions correspond to different 2x2 contingency tables. We seek conditions that will best distinguish between the models of interest. Several details complicate this search. The first is that psychological theories of causal inference are often intended only as ordinal descriptions of judgment. Cheng (1997), for instance, proposes that power PC theory predicts only people's rankings of causal strengths. That is, the model asserts $w_1 < w_2$ implies only that $J(w_1) < J(w_2)$, where w_i is the causal power of cause C_i and $J(.)$ gives the subjective rating for the causal powers. One difficulty, then, is that two models may differ quantitatively while giving ordinal predictions that are highly similar or identical.

Another difficulty arises when model predictions depend on free parameters. The weighted ΔP model depends on the free parameter θ. Consequently, the predicted ranking of causal strengths will be somewhat contingent on this parameter.

Good empirical comparisons will be resilient to the aforementioned complications. Namely, they will rely on predictions that are the least sensitive to distortions of subjective evaluation so that $w_i - J(w_i)$ differences are likely to be small. In addition, good comparisons are ones that are not strongly contingent on the choice of free parameters.

Let the vector [a,b,c,d] correspond to the four entries of a 2x2 contingency table and assume they are all positive unless explicitly noted otherwise. The generative weighted ΔP model will not depend on the free parameter for any conditions with $c = 0$, or equivalently, with base rate $P(e^+|c^-) = 0$. This is evident from the expression:

$$\begin{aligned} w\Delta P &= P(e^+|c^+) - (1-\theta)P(e^+|c^-) \\ &= P(e^+|c^+) \end{aligned}$$

A good place to start, then, are conditions in which the base rate of the effect equals zero. In these conditions weighted ΔP predicts judgments equal to $P(e^+|c^+)$. Similarly, for preventive weighted ΔP the model is parameter free when $d = 0$ or $P(e^-|c^-) = 0$.

There is still the issue of probability representation. Within the cumulative prospect theory literature, four families of functions have been proposed for the subjective representation of probabilities (Cavagnaro, Pitt, Gonzalez, & Myung, 2013). The general consensus is that the probability weighting function $W(p)$ has an inverse sigmoid shape and that objective values are returned at the endpoints so that $W(0) = 0$ and $W(1) = 1$. If similar weighting applies to the causal learning task, then subjective representation should match objective probability for deterministic events. This seems reasonable, as it simply requires that people assign a strength of 0 when they think the event will never occur and a strength of 1 when they think the event will always occur.

Generative weighted ΔP always predicts deterministic strength judgments of 0 for conditions with [0,b,0,d] contingencies and judgments of 1 for [a,0,0,d] contingencies. This is because the free parameter θ does not influence these predictions. Further, these judgments are unlikely to be distorted by subjective representation. In sum, the weighted ΔP makes unambiguous predictions for this pair of conditions.

In contrast, any reasonable Bayesian model will predict probabilistic strengths greater than 0 for the [0,b,0,d] conditions and less than 1 for the [a,0,0,d] conditions (what is meant by "reasonable" is made explicit in Appendix F). Thus, weighted ΔP and Bayesian models give predictions that differ in kind. Comparisons over these conditions should therefore constitute a strong test of weighted ΔP against Bayesian models. Finally, a parallel argument can be made for using the [a,0,c,0] and [0,b,0,d] contingencies when contrasting predictions for preventive causes.

4.3 Aggregate measures of causal judgment

The standard approach in causal learning studies is to report sample means from each condition, and then fit the competing models to these means. Though this approach is the norm, it does complicate inference to the individual level. Models of causal judgment are generally intended to describe individual level behavior. But as Danks and Eberhardt (2011) observe, if one uses an aggregate measure, such as the mean, then it should be compared to the corresponding aggregate of the model predictions. Aggregate predictions require a more elaborate model that includes an account of error or individual variation. The original formulation of weighted ΔP and Bayesian models comes with no such account. Without a description of error, the best that can be done is to choose an aggregate measure that well-represents the "typical" individual response for each condition. One goal of this chapter is to evaluate whether group means, or some other aggregate measure, is a reasonable proxy for typical individual responses.

4.4 Experiment 1

The experiments described below are quite similar to those performed by Collins and Shanks (2006). In these experiments, participants read a novel cover story in which they are asked to interpret hypothetical laboratory results. Novel cover stories are used to ensure that participants all have the same hypothesis space and the same prior beliefs about the task domain (Danks & Eberhardt, 2011). It should also lead them to assign roughly equal utility to the different hypotheses, and so participants should primarily be concerned with accuracy when formulating their predictions.

The goal of experiment 1 was to collect individual-level data to be used in the critical comparisons discussed above. In addition, response distributions are used to evaluate various aggregate measures of performance.

4.4.1 Methods

Participants. Undergraduate students from the University of Washington ($N = 285$) participated in the task. All participants were recruited through introductory psychology courses and awarded a small amount of course credit. Participants were 58% female and ages ranged from 16 to 25.

Design and Procedure. Participants enrolled through the psychology subject pool program, where they were provided a link to the study. The experiment was administered via the internet using the Qualtrics software package. Each participant worked through one practice condition before random assignment to 2 out of the 15 experimental conditions. The contingencies for each condition are shown in Table 4.1 below. Participants were given the following instructions:

> During this task you will see laboratory records from two studies. In each study, you will see information about administering a particular protein to a different species of butterfly. In a test given some time later, the butterflies were examined for whether their *OPTIX* gene was turned on.
>
> Of course, regular cell processes can cause activation of the *OPTIX* gene even in insects that are not given the protein treatment. What you must decide is whether and how strongly the proteins administered to the butterflies in the experiment can independently cause gene activation.
>
> There are 32 butterflies in each study. Half of the butterflies in each study were randomly assigned to a group receiving the protein, and half to a group not receiving the protein. Each record tells you whether or not a particular butterfly has been exposed to the relevant protein, and you will be asked to predict whether a test given later will find that this butterfly's *OPTIX* gene has been turned on.
>
> When you have made your prediction you will be told if gene activation occurred. Use this feedback to try to find out whether the protein really causes gene activation. Although initially you will have to guess, by the end you will be an expert.
>
> At regular intervals during each study you will be asked to estimate the degree to which the protein causes gene activation, and to state how confident you are in your estimate. Further instructions will explain at the appropriate time how to make these estimates. Please try to be as accurate as possible. You can now try some practice trials before the main test begins.

Learning trials were presented sequentially in blocks of sixteen, with trial order randomized within each block. Each trial showed a color picture of the butterfly and whether it was treated with the protein. Participants were asked to use the keyboard to predict whether “yes, the gene will be turned ON” or “no, the gene will be turned OFF”. Immediately after the prediction they were given feedback that showed a color coded image of a gene that was also labeled as either “ACTIVE” or “INACTIVE”. After 16 trials participants were asked to make a causal strength estimate. The participant then worked through another 16 trials which were followed with a four questions. They were asked to make a final causal strength judgment. A variant of the standard causal probe was used for the strength estimate by asking participants:

> Please estimate the degree to which the protein causes gene activation.
>
> Select a value between 0 and 100, where 0 indicates that the protein never causes activation and 100 indicates that the protein always causes gene activation. Intermediate numbers indicate intermediate levels of causal influence.

Below the question was a slider scale with tick marks labeled at 10-unit increments, though participants were free to select any number between 0 and 100. The beginning position of the slider was at a rating of 50. Above the 0 on the scale at the far left was text that said "Never Activates" and above the 100 on the right was the text "Always Activates". To the left of the scale was the label "Strength of Protein Influence". Participants made their judgments by using the mouse to click on the scale. On the mouse click, a number appeared to the right of the scale, which indicated the value of the judgment.

Immediately below the causal probe was a question asking, "How confident are you in the judgment you have just made on a scale from 0 (not at all confident) to 100 (certain)?". This question was also accompanied by a 0 to 100 slider with the label "Uncertain" over the 0 tick mark and the label "Certain" over the 100 tick mark.

After submitting their causal strength and confidence judgments participants were presented with two additional questions that asked them to estimate the conditional probabilities $P(e^+|c^+)$ and $P(e^+|c^-)$. Specifically, for $P(e^+|c^+)$ they were asked:

> Suppose you observe another collection of ***protein-treated*** *Gonepteryx Formosana* butterflies. Please estimate the percentage that will have activated genes.

And for $P(e^+|c^-)$ they were asked:

> Suppose you observe another collection of ***untreated*** *Gonepteryx Formosana* butterflies. Please estimate the percentage that will have activated genes.

Pictures of the butterfly with the cause present or absent accompanied each of the questions. Once again, participants used a 0 to 100 slider scale to make their judgments. The order of these questions was counterbalanced across participants.

4.4.2 Results

I begin by examining responses for the key conditions identified in Section 4.2. For the [0,8,0,8] condition, 35 out of 42 participants gave strength ratings of 0. And for the [8,0,0,8] condition, 32

out of 40 participants gave strength ratings of 100. Thus, in both conditions a large majority of respondents returned the point prediction of the weighted ΔP model.

Next, I explore causal strength response distributions within each condition, as shown in Figure 4.1. Several conditions have strongly skewed distributions with modal judgments clustered at 0 or 100. The best examples are the just mentioned [0,8,0,8] and [8,0,0,8] conditions. However, other experimental conditions also demonstrate these features. In particular, conditions with observed $P(e^+|c^+) = 1$ are all left-skewed with many responses of 100. Consequently, median judgments are reported since they appear to better reflect typical responses. The decision to use medians is justified more fully in the discussion below.

In order to use and compare medians it is necessary to characterize their variance. Wilcox (2010, 2012) observes that a high proportion of tied scores can create problems when attempting to compute confidence intervals for the median. He goes on to recommend *percentile bootstrap confidence intervals*, as they have been shown to perform well even with many tied ranks in the data. Since tied ranks are quite common in some of the experimental conditions, such as the two just discussed above, I follow Wilcox's advice and compute 95% percentile bootstrap confidence

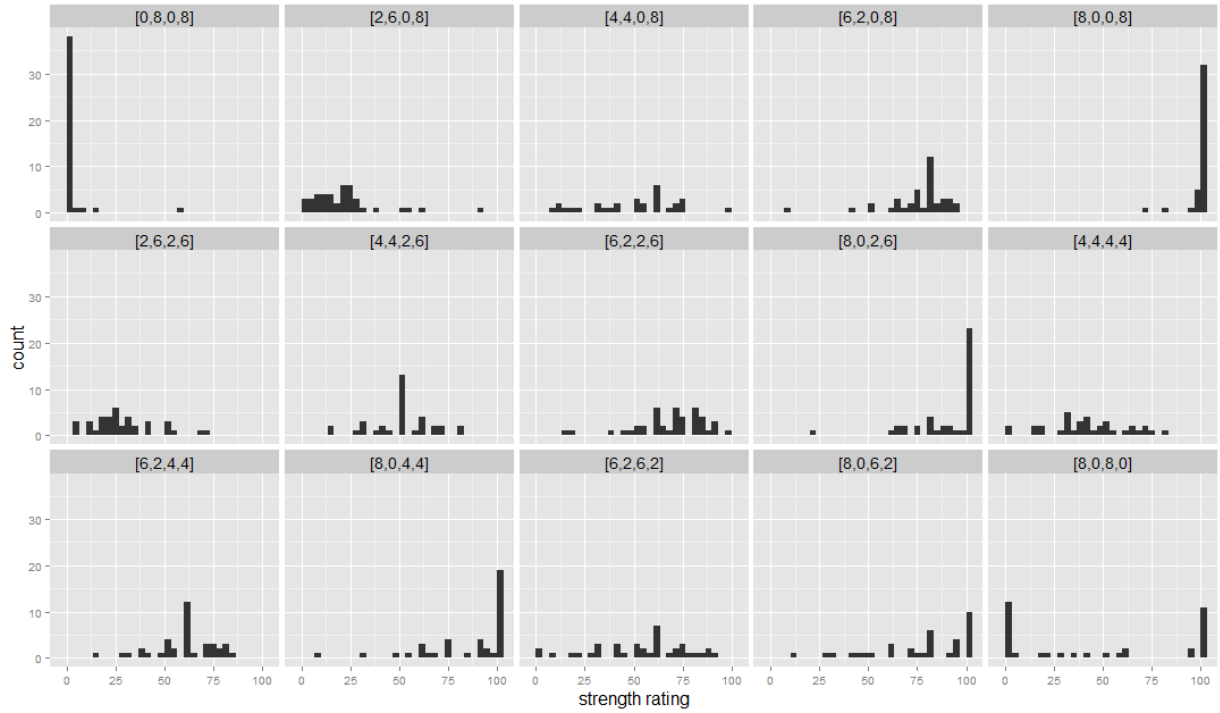


Figure 4.1. Histograms of response counts across the 15 conditions of experiment 1. Condition labels of [a,b,c,d] give the four frequencies of the 2x2 contingency table.

intervals for each condition. Specifically, I use the binomial method of finding approximate confidence intervals (cf. Wilcox, 2012, p. 130).

Median judgments with 95% bootstrap confidence intervals and model predictions are shown in Table 4.1 and plotted in Figure 4.2. Notice that the conditions in the figure are grouped by the base rate of the effect. Overall, the weighted ΔP model does quite well describing the data, with 12 out of 15 predictions within the 95% confidence intervals and two of the misses occurring right on the boundary. Note that in the key [0,8,0,8] and [8,0,0,8] conditions, the 95% bootstrap intervals have no length. This results from the strong modal responses of 0 and 100 respectively in these two conditions.

Table 4.1. Design and results of experiment 1: generative component

Experiment	*a*	*b*	*c*	*d*	median rating	bootstrap 95% CI	MLE	$w\Delta P$	uniform Bayes
	0	8	0	8	0	[0,0]	0	0	10
	2	6	0	8	20	[13,24]	25	25	25
	4	4	0	8	54	[39,60]	50	50	44
	6	2	0	8	80	[75,81]	75	75	66
	8	0	0	8	100	[100,100]	100	100	88
	2	6	2	6	25	[20,30]	0	18	20
	4	4	2	6	50	[50,52]	33	43	34
	6	2	2	6	72	[61,76]	67	68	56
	8	0	2	6	100	[89,100]	100	93	84
	4	4	4	4	40	[33,49]	0	36	27
	6	2	4	4	60	[56,63]	50	61	45
	8	0	4	4	97	[85,100]	100	86	77
	6	2	6	2	60	[45,62]	0	54	37
	8	0	6	2	80	[73,95]	100	79	67
	8	0	8	0	40	[0,95]	NA	71	56

Notes. The weighted ΔP free parameter $\theta = .71$ gave best fit to human median judgments across both generative and preventive conditions (see Table 4.2) as measured by mean squared error.

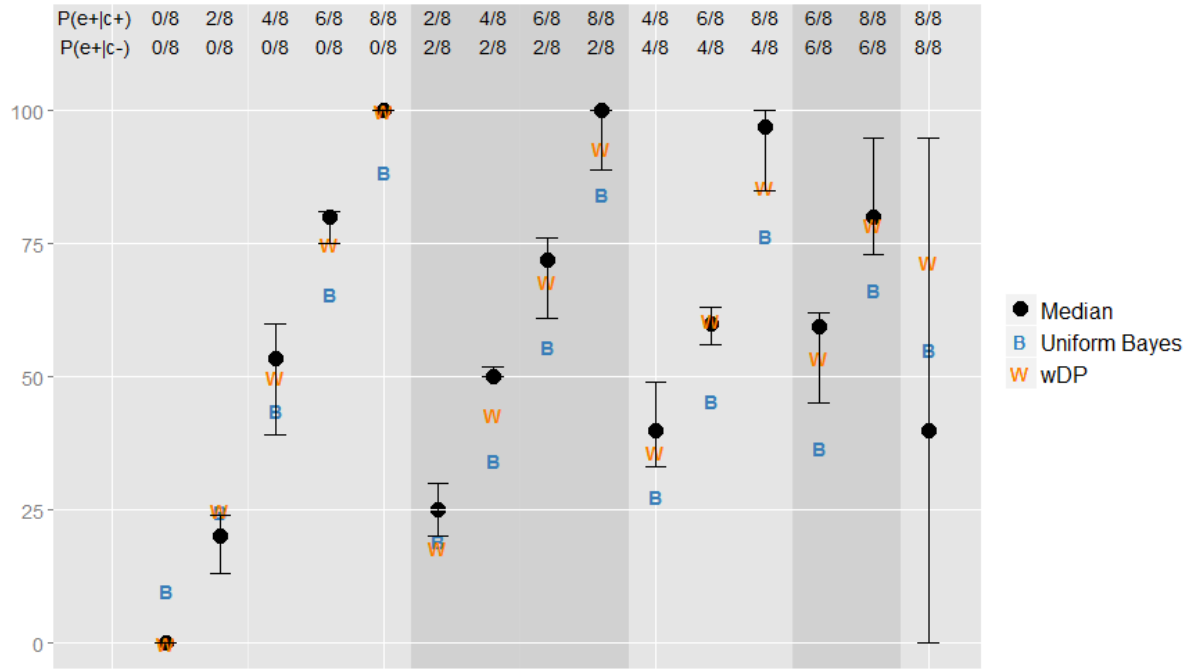


Figure 4.2. Median judgments, 95% bootstrap confidence intervals, and model predictions across the 15 conditions of experiment 1. Conditions are grouped by the observed base rate of the effect $P(e^+|c^-)$.

Condition [8,0,8,0] has an extremely wide confidence interval, stretching from 0 to 95. This is due to the bimodality of the response distribution (Figure 4.1). Of the 35 respondents, 12 gave a rating of 0, 11 gave a rating of 100, and the remaining 12 were distributed across the interior of the interval. Hence, there appear to be several distinct response strategies within this condition, a topic considered in more depth below.

Now focus on the first five conditions in which the observed base rate of the effect is equal to 0. For these conditions the weighted ΔP model has no free parameters. Thus, for these conditions *all* weighted ΔP models predict strength ratings equal to $P(e^+|c^+)$. These predictions are generally borne out, with four of five within the 95% confidence interval. The lone miss occurs for the [2,6,0,8] condition in which the weighted ΔP prediction is 25 while the 95% interval extends from 13 to 24.

Since participants were also asked to judge $P(e^+|c^+)$, it is possible to directly test the weighted ΔP predictions for the zero base-rate conditions. Specifically, we can test people's causal strength rating against their subjective probability assessments for $P(e^+|c^+)$. The percentile bootstrap method can again be used by randomly sampling pairs of observations from the data.

Table 4.2. Percentile bootstrap 95% confidence intervals for median difference scores of causal strength ratings minus $P(e^+|c^+)$ ratings.

Experiment	*a*	*b*	*c*	*d*	median difference score	Bootstrap 95% CI
	0	8	0	8	0	[0,0]
	2	8	0	8	0	[-6,0]
	4	8	0	8	-6.5	[-9,0]
	6	8	0	8	0	[-1,0]
	8	8	0	8	0	[0,0]

Medians of the difference scores and 95% bootstrap confidence intervals are shown in Table 4.2. The results strongly support the weighted ΔP prediction that, when the base rate of the effect $P(e^+|c^-)$ is zero, causal strength ratings equal the judged probability for $P(e^+|c^+)$. The median difference score equals exactly 0 in 4 of the 5 conditions. There is a small nonzero median difference for the [4,8,0,8] condition, though the associated 95% confidence does include 0, implying that the hypothesis of "no difference" remains plausible.

4.4.3 Discussion

Exploratory analysis revealed that participants are predisposed to give ratings of 0 or 100 for a number of the experimental conditions. Henceforth, I refer to this predisposition as the *deterministic bias*. I refer to 0 and 100 as *deterministic ratings* since they indicate that the effect does not or does occur with certainty. Correspondingly, ratings from 1 to 99 will be referred to as *probabilistic ratings*.

The deterministic bias was strongest in the [0,8,0,8] and [8,0,0,8] conditions, for which a large majority gave ratings of 0 and 100, respectively. This finding provides strong evidence against *all* Bayesian models of causal power. It can be shown (Appendix F) that no Bayesian model of parameter estimation will return the 0 and 100 predictions for the [0,8,0,8] and [8,0,0,8] conditions and also give probabilistic predictions for the other conditions. Typical Bayesian models, including the uniform and SS prior models, will give probabilistic predictions for the [0,8,0,8] and [8,0,0,8] conditions. Nonetheless, a minority of participants did give small nonzero strength ratings in the

[0,8,0,8] condition. Similarly, in the [8,0,0,8] condition a small group gave high ratings that were below 100. So at least some respond probabilistically in these conditions. While these ratings are more consistent with a Bayesian model, much additional evidence is necessary to show that this subgroup is in fact using something like Bayesian inference.

The deterministic bias also appears in a number of other conditions. A large proportion of 100 ratings occur in the [8,0,2,6], [8,0,4,4] and [8,0,6,2] conditions (Figure 4.1). Meanwhile, the [8,0,8,0] condition is strongly heterogeneous with about a third of ratings equal to 0, a third equal to 100, and the remaining third probabilistic. This heterogeneous pattern leads to an extremely wide 95% confidence interval (Figure 4.2) stretching from 0 to 95.

Thus, for a number of conditions there appears to be a mixture of response strategies with some respondents giving deterministic 0 or 100 ratings while others give probabilistic ratings. As point-estimators of causal strength, the weighted ΔP and Bayesian models cannot account for such mixed response patterns. I take up the general issue of mixed deterministic and probabilistic responses in the general discussion below.

The use of group means, used in previous causal learning studies, will conceal any potential deterministic bias. Absent strict consensus, the mean rating for a condition will always be probabilistic as it will occur between 0 and 100. Indeed, of the 114 conditions used in the Perales and Shanks (2007) meta-analysis, none of them had aggregate scores equal to 0 or 100. Yet given the response distributions shown in Figure 4.1, there is good reason to believe that the mean is not a good representation of individual strategies. Group means will not be representative of a typical response when distributions are highly skewed or if responses fall into several distinct clusters. The median, in contrast, will equal 0 or 100 so long as a majority of individuals give these ratings in a given condition. In such cases, the median is identical to the mode, and so it represents the rating given by most participants.

The median is often a preferred aggregate measure when observations are drawn from highly skewed distributions. This is why it is routinely used to characterize income or wealth distributions. In conditions for which there is a relatively symmetric response pattern, mean and median judgments will be quite similar, and so there is no strong reason to prefer one over the other.

Given the properties of the data, the median seems a better aggregate measure of a “typical” response for a given condition. Nonetheless, when several distinct response strategies are present,

the best an aggregate summary can do is to represent the dominant strategy. And when there is no dominant strategy, such as with the [8,0,8,0] condition, no aggregate point measure will be representative of a typical response.

4.5 Experiment 2

The purpose of experiment 2 was to see if the experiment 1 findings also hold for preventive causes. The data from experiments 1 and 2 are also used to evaluate competing linear combination models of causal judgment, as described in a later section. Experiment 2 is almost identical to experiment 1, except now participants are presented with a scenario in which the cause potentially prevents the effect.

4.5.1 Methods

Participants. Undergraduate students from the University of Washington ($N = 325$) participated in the task. All participants were recruited through introductory psychology courses and awarded a small amount of course credit. Participants were 59% female and ages ranged from 15 to 35.

Design and Procedure. Design and procedure are identical to experiment 1 except where noted below. The experimental contingencies, shown in Table 4.3, now reflect the action of a preventive cause so that the effect tends to occur less often with the cause present. The task instructions were also slightly modified to indicate that the potential cause was preventive. The instructions read:

> Imagine you are working in a laboratory studying several species of threatened butterflies. The widespread use of pesticides has presented a serious problem for these populations. Previous research has shown that some protection against pesticides occurs when a butterfly's OPTIX gene is turned off.
>
> Your lab has begun to work on protein treatments that you hope will prevent activation of the OPTIX gene in the butterfly genome. In recent pilot studies you have administered these proteins to several of the threatened species.
>
> During this task you will see laboratory records from two studies. In each study, you will see information about administering a particular protein to a different species of butterfly. In a test given some time later, the butterflies are examined for whether their OPTIX gene was turned on.
>
> There are 32 butterflies in each study. Half of the butterflies in each study were randomly assigned to a group receiving the protein, and half to a group not receiving the protein. Each record tells you whether or not a particular butterfly has been exposed to the relevant protein, and you will be asked to predict whether a test given later will find that this butterfly's OPTIX gene has been turned on.

> When you have made your prediction you will be told if gene activation occurred. Use this feedback to try to find out whether the protein really prevents gene activation. Although initially you will have to guess, by the end you will be an expert.
>
> At regular intervals during each study you will be asked to estimate the degree to which the protein prevents gene activation, and to state how confident you are in your estimate. Further instructions will explain at the appropriate time how to make these estimates. Please try to be as accurate as possible.
>
> You can now try some practice trials before the main test begins.

Learning trials were presented as in experiment 1 with participants asked to predict whether the gene was on or off. After 16 trials, participants were asked to make a causal strength estimate. Then after another 16 trials they were asked to make a final causal strength judgment, give a confidence rating for the judgment, and also estimate $P(e^+|c^+)$ and $P(e^+|c^-)$. The standard causal probe used for the strength judgment was:

> Please estimate the degree to which the protein prevents gene activation.
>
> Select a value between 0 and 100, where 0 indicates that the protein does not prevent activation and 100 indicates that the protein always prevents gene activation. Intermediate numbers indicate intermediate levels of preventive influence.

Participants again used a slider scale to make their causal strength ratings. Now above the 0 on the scale at the far left was text that said "Does not prevent" and above the 100 on the right was the text "Always Prevents". The confidence and conditional probability questions were identical to those described in experiment 1.

4.5.2 Results

Results are presented in the same order as was done for the generative component, beginning with the key comparisons discussed at the beginning of the chapter. Recall that preventive weighted ΔP is defined with respect to $P(e^-|c^+)$ and $P(e^-|c^-)$. So now the weighted ΔP model gives a strength of 0 for the [8,0,8,0] condition and a strength of 1 for the [0,8,8,0] condition. Moreover, preventive weighted ΔP does not depend on the free parameter θ whenever $P(e^-|c^-) = 0$, or for conditions in which the effect always occurs when the cause is absent since $P(e^+|c^-) = 1 - P(e^-|c^-) = 1$.

In the [8,0,8,0] condition, 24 out of 40 participants gave a strength rating of 0. And for the [0,8,8,0] condition, 32 out of 49 participants gave strength ratings of 100. Once again, in both

conditions a majority of respondents conformed to the point prediction of the weighted ΔP model. However, the majorities were smaller in comparison to the experiment 1. I speculate on what may be driving this difference in the discussion below.

The causal strength response distributions of each condition are shown in Figure 4.3. The conditions in Figure 4.3 are arranged in the same order as Figure 4.1 by ΔP and causal power. Equivalently, the conditions arranged by $P(e^+|c^+)$ and $P(e^+|c^-)$ in Figure 4.1 are correspondingly arranged by $P(e^-|c^+)$ and $P(e^-|c^-)$ in Figure 4.3. The corresponding conditions across the two figures have similar response distributions, though the preventive conditions appear to have somewhat more dispersion. As before, a number of conditions are skewed with modal judgments clustered at 0 or 100. In Figure 4.3, conditions with observed $P(e^-|c^+) = 1$ are all left-skewed with many responses of 100.

Median judgments with 95% bootstrap confidence intervals and model predictions are shown in Table 4.3 and plotted in Figure 4.4. Conditions are grouped by the base rate of the effect. Weighted ΔP does a good job describing the data, with 11 out of 15 predictions within the 95% confidence intervals and several misses occurring right on the boundary. Once again, the key

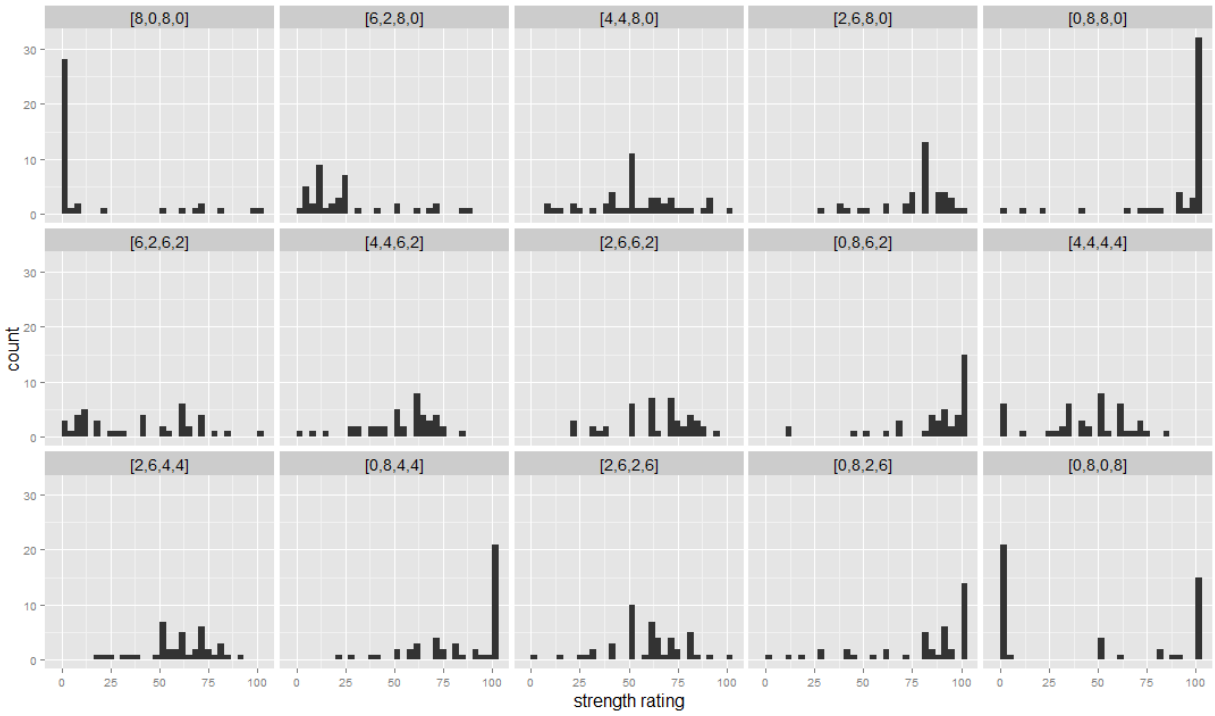


Figure 4.3. Histograms of response counts across the 15 conditions of experiment 2. Condition labels of [a,b,c,d] give the four frequencies of the 2x2 contingency table.

conditions have extremely narrow confidence intervals. The [8,0,8,0] condition has a 95% confidence interval from 0 to 1 while the [0,8,8,0] condition has no length centered on 100. This is due to the respective modal responses of 0 and 100 in these two conditions.

Condition [0,8,0,8] has an extremely wide confidence interval from 0 to 88, which is due to a bimodality in the response distribution (Figure 4.3). Of the 46 respondents, 19 gave a rating of 0, 15 gave a rating of 100, and the remaining 12 were distributed across the interval. Interestingly, four respondents gave a rating of exactly 50, which is sometimes used as a proxy for "uncertain". Overall, the ratings closely parallel the result for the [8,0,8,0] condition in the generative experiment.

Table 4.3. Design and results of experiment 2: preventive component

Experiment	*a*	*b*	*c*	*d*	median rating	bootstrap 95% CI	MLE	$w\Delta P$	uniform Bayes
	8	0	8	0	0	[0,1]	0	0	10
	6	2	8	0	19	[12,25]	25	25	25
	4	4	8	0	50	[50,60]	50	50	44
	2	6	8	0	80	[75,83]	75	75	65
	0	8	8	0	100	[100,100]	100	100	88
	6	2	6	2	40	[18,60]	0	18	19
	4	4	6	2	60	[50,61]	33	43	34
	2	6	6	2	68	[60,71]	67	68	56
	0	8	6	2	94	[87,99]	100	93	84
	4	4	4	4	48	[36,50]	0	36	27
	2	6	4	4	60	[52,69]	50	61	46
	0	8	4	4	94	[74,100]	100	86	77
	2	6	2	6	61	[50,66]	0	54	36
	0	8	2	6	90	[80,94]	100	79	67
	0	8	0	8	50	[0,88]	NA	71	54

Notes. The weighted ΔP free parameter $\theta = .71$ gave best fit to human median judgments across both generative (see Table 4.1) and preventive conditions as measured by mean squared error.

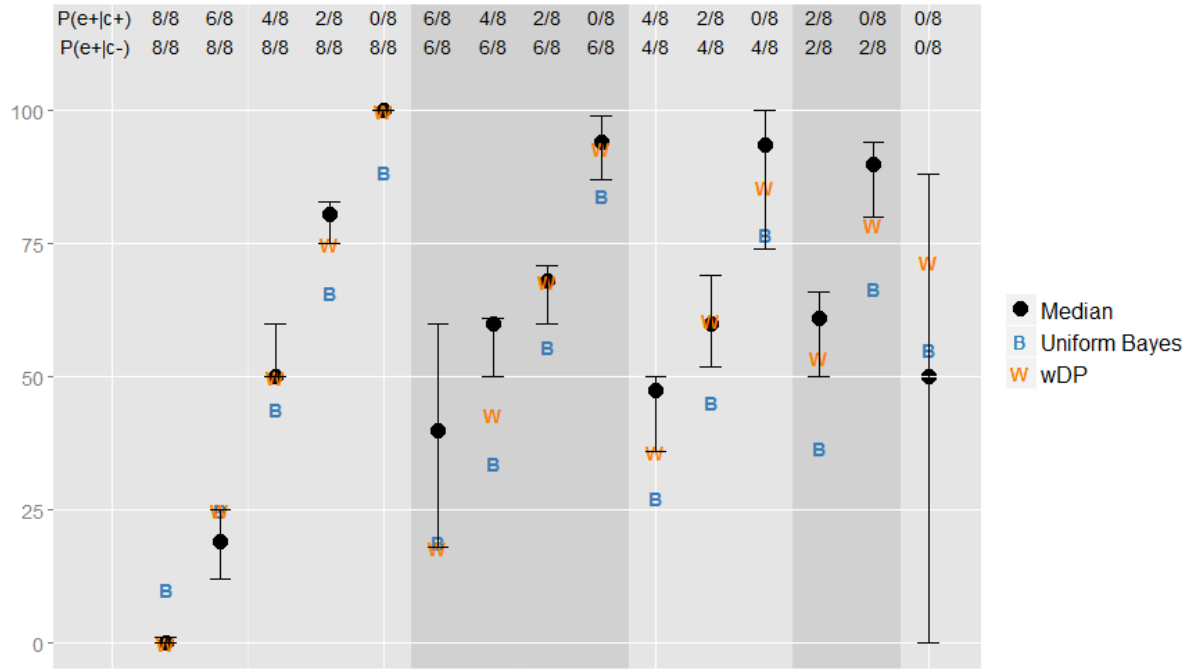


Figure 4.4. Median judgments, 95% bootstrap confidence intervals, and model predictions across the 15 conditions of experiment 2. Conditions are grouped by the observed base rate of the effect $P(e^+|c^-)$.

The first five conditions in Table 4.3 and Figure 4.4 have the observed base rate of the effect equal to 1. In these conditions, the weighted ΔP model has no free parameters since the weight multiplies $P(e^-|c^-) = 0$. Again, for these conditions *all* weighted ΔP models predict strength ratings equal to $P(e^-|c^+)$. Observed judgments agree with weighted ΔP, with all five of the 95% confidence intervals containing the predicted values.

Once again, participant judgments for $P(e^+|c^+)$ may be used to test predictions for the $P(e^+|c^-) = 1$ conditions. Preventive weighted ΔP predicts judgments of $P(e^-|c^+)$ for these conditions. Hence, causal strength ratings are compared to 100 minus the subjective probability assessments for $P(e^+|c^+)$ since $P(e^-|c^+) = 1 - P(e^+|c^+)$. Medians of the difference scores and 95% bootstrap confidence intervals are shown in Table 4.4. Again, the data strongly support the weighted ΔP prediction. Median difference scores are exactly 0 in 4 out of the 5 conditions. There is a small nonzero difference for the [2,6,8,0] condition, though the associated 95% confidence interval does cover 0.

Table 4.4. Percentile bootstrap 95% confidence intervals for median difference scores of causal strength ratings minus $[100 - P(e^+|c^+)]$ ratings.

Experiment	*a*	*b*	*c*	*d*	median difference score	Bootstrap 95% CI
	8	0	8	0	0	[0,0]
	6	2	8	0	0	[-9,0]
	4	4	8	0	0	[0,4]
	2	6	8	0	2	[0,20]
	0	8	8	0	0	[0,1]

4.5.3 Discussion

The results from experiment 2 replicate the findings from experiment 1. A deterministic bias was observed in conditions [8,0,8,0] and [0,8,8,0] in which a majority of people gave ratings of 0 and 100, respectively. Once again, this finding is incompatible with *all* Bayesian models of strength estimation. The deterministic bias was also present in a number of additional conditions. Whenever the effect failed to occur, some proportion of respondents gave causal strength ratings of 100. And in the [0,8,0,8] condition there was a bimodal response with many strength ratings of 0 or 100. As in experiment 1, there is evidence for a mixture of causal judgment strategies.

One potentially interesting difference between experiments 1 and 2 is that there appears to be more variation in responses for the preventive conditions. There is a median confidence interval width of 13 for the preventive experiment versus a median width of 11 across the generative conditions. Moreover, the preventive confidence intervals are wider despite more observations per condition relative to the generative experiment. A possible explanation is that participants find the preventive conditions more difficult. Such an explanation would seem to agree with the mechanics of the preventive weighted ΔP model. Recall that under the preventive model, a "success" must be coded as the *absence* of an occurrence. Coding occurrences as "successes" would seem to be more natural than coding a non-occurrence as such. If this is in fact true, then preventive weighted ΔP will require more cognitive effort and should lead to more errors.

4.6 Cross-validation study

The predictions presented above are from a weighted ΔP model with a single free parameter θ that represents prior expected strength for both generative and preventive causes. Specifically, they are from the 1-parameter weighted ΔP model:

$$\widehat{w}_1 = \begin{cases} P(e^+|c^+) - (1-\theta)\times P(e^+|c^-) & \text{for generative } C \\ P(e^-|c^+) - (1-\theta)\times P(e^-|c^-) & \text{for preventive } C \end{cases} \quad (4.1)$$

Of course, there are other possible linear combination models. For instance, prior expectations may differ across generative and preventive contexts. This would result in the more general model:

$$\widehat{w}_1 = \begin{cases} P(e^+|c^+) - \left(1-\theta_g\right)\times P(e^+|c^-) & \text{for generative } C \\ P(e^-|c^+) - \left(1-\theta_p\right)\times P(e^-|c^-) & \text{for preventive } C \end{cases} \quad (4.2)$$

where θ_g and θ_p are prior expected causal powers for respective generative and preventive contexts. Thus, model (4.1) is a special case of (4.2) with the restriction $\theta_g = \theta_p = \theta$. Finally, in Chapter 3 we saw the most general linear combination model:

$$\widehat{w}_1 = \begin{cases} \gamma_0 + \gamma_1 P(e^+|c^+) + \gamma_2 P(e^+|c^-) & \text{for generative } C \\ \pi_0 + \pi_1 P(e^+|c^+) + \pi_2 P(e^+|c^-) & \text{for preventive } C \end{cases} \quad (4.3)$$

where $\{\gamma_0, \gamma_1, \gamma_2\}$ are the weights for generative contexts and $\{\pi_0, \pi_1, \pi_2\}$ are the weights for preventive contexts. We also saw in Chapter 3 how (4.3) is the more general case of (4.2). The generative model is given by the restrictions $\gamma_0 = 0$, $\gamma_1 = 1$ and $\gamma_2 = -\left(1-\theta_g\right)$ while the preventive model is from the restrictions $\pi_0 = \theta_p$, $\pi_1 = -1$ and $\pi_2 = \left(1-\theta_p\right)$.

The choice of (4.1) as the preferred model can be justified, in part, by its strong performance in the replication of Perales and Shanks (2007) cross-validation study. However, the data collected in experiments 1 and 2 above can also be used to create our own model competition study. Indeed, access to individual observations allows for a cross-validation approach that is better suited to the model selection task. In the next section I describe this cross-validation method. Then I apply the method to assess models (4.1) through (4.3) to see which one gives the best account of the data.

4.6.1 Method

Cross-validation is the standard method used to estimate a model's prediction or test error (Hastie et al., 2009; James, Witten, Hastie, & Tibshirani, 2015). Recall that Perales and Shanks (2007) began with mean judgments from 114 experimental conditions. For each calibration-validation iteration, they randomly split the sample of conditions, fit the models to half of the condition means, and evaluated predictions against the other half. On this approach, conditions are randomly sampled while judgments for a condition are treated as fixed. The approach, then, tells us how well a model calibrated to one set of experiments will generalize to a new set of different experiments. To the extent that there is redundancy among experimental conditions, their approach will also inform how well model predictions generalize from one set of participants to another.

Ideally, we want to know a model's predictive accuracy over the greatest diversity of contexts. This will ensure that the model is strong over the most general set of circumstances. Perales and Shanks were limited by only having group means for each condition, and so they could only fit and test models against a subset of conditions in their cross-validation study. Access to the individual observations within each condition frees us from this constraint. The below cross-validation study randomly samples observations *within* each condition. This allows for the use of the entire set of experimental conditions on each iteration. Specifically, on each iteration observations are randomly divided into calibration and validation samples within each condition. Summary statistics, such as medians, are found for each condition using the calibration and validation samples. These medians can then be used to fit and evaluate the competing models.

On this approach, then, the set of conditions are fixed while individual observations are sampled. The approach informs us how well models will generalize to new sets of observations gathered in the same experimental conditions. By using every condition, the approach maximizes the diversity of contexts against which predictions are evaluated. Thus, it is paramount that the conditions employed constitute a balanced sample from the space of all possible experiments. The conditions of experiments 1 and 2 constitute the same set as those used in Wasserman et al. (1993) and Buehner and Cheng (1997). These studies were each intended as comprehensive assessments of causal learning. So it is reasonable to believe that experiments 1 and 2 form a balanced sample from the space of all possible contingency tables.

An additional amendment concerns the formation of the calibration and validation samples. As mentioned above, Perales and Shanks used a 50-50 split in forming these samples. A disadvantage of this method is that it throws away half of the sample for the training set, which may lead to an overestimate of the prediction error (Hastie et al., 2009; James et al., 2015). This risk will be exaggerated if the sample size is moderate to small and the fitted model has many free parameters.

One alternative to the 50-50 split is leave-one-out cross-validation. On this method, models are fit to the entire sample except for a single observation. The difference between the model prediction and the single observation can be used as an estimate of the prediction error, and this process can be repeated for every observation in the data set. The advantage of leave-one-out cross-validation is that it is a less biased estimate of prediction error. Yet the bias-variance trade-off also holds for the estimation of prediction error. While leave-one-out cross-validation is low bias, it is also very high in variance.

In lieu of a 50-50 split or leave-one-out cross-validation, researchers now advocate for an intermediate method known as *k*-fold cross-validation (Hastie et al., 2009; James et al., 2015). The mechanics are quite similar to the previous approaches. First, observations are randomly divided into *k*-groups or folds. The first group is held out as a validation sample while the model is fit to the data in the remaining $k - 1$ groups. Model predictions are then compared to the first group in order to estimate test error. The process is then repeated, with each of the *k* groups serving as the validation set and the remaining $k - 1$ groups serving as the calibration set. Statistical learning researchers have found that using $k = 5$ or $k = 10$ tends to have good properties in terms of balancing bias and variance in the estimation of prediction error.

In summary, this cross-validation study randomly divides observations within each condition, thus allowing all conditions from experiments 1 and 2 to be used on each iteration. The study uses *k*-fold cross validation with both $k = 5$ and $k = 10$. Test error is estimated using the validation sample mean squared error, averaged over each of the *k* folds. Finally, the *k*-fold procedure is iterated to ensure robustness of the results. There are 2000 iterations for $k = 5$ while there are 1000 iterations for $k = 10$. This gives 10,000 estimates of the prediction error for each of the simulations.

4.6.2 Results

Median mean-squared errors and median parameter estimates from the cross-validation study are shown in Table 4.5. The 1-parameter model corresponds to (4.1), the 2-parameter model to (4.2), and the linear combination model to (4.3), above.

Across both the $k = 5$ and $k = 10$ folds, the 1-parameter model has the lowest median MSE, followed by the linear combination model and then the 2-parameter model. Overall, MSEs for the 1-parameter and linear combination model are quite close. Parameter estimates are approximately the same across the two fold sizes. Note that the median parameter estimates for the linear combination model are nearly equal to the restrictions implied by the weighted ΔP model.

MSE for each model is higher in the $k = 10$ than in the $k = 5$ simulation. The stability of the parameter estimates suggests an explanation for this result: If parameter estimates are relatively unchanged from $k = 5$ to $k = 10$, then the primary driver of MSE will be sampling variability in the validation sample. And since $k = 10$ has smaller validation samples it will result in a larger average MSE.

Though not reported in the table, the study also evaluated was Perales and Shanks' (2007) EI rule. Median fit was quite poor with an MSE of 982 for $k = 5$ and an MSE of 1026 for $k = 10$. One can also specify a directed EI rule that allows cell weights to vary with causal direction, similar to the linear combination model above. The directed EI rule had much better median fit, though was not better than the models reported in Table 4.5.

Table 4.5. Median MSE of prediction and median parameter estimates for the *k*-fold cross-validation study.

	k=5		*k=10*	
Model	Median MSE	Median parameter fits	Median MSE	Median parameter fits
1-parameter	243	$\theta = .70$	287	$\theta = .71$
2-parameter	260	$\theta_g = .68$ $\theta_p = .74$	300	$\theta_g = .67$ $\theta_p = .75$
Linear combination	250	$\{\gamma_0 = .04, \gamma_1 = 1.03, \gamma_2 = -.42\}$ $\{\pi_0 = .71, \pi_1 = -.93, \pi_2 = .31\}$	290	$\{\gamma_0 = .04, \gamma_1 = 1.03, \gamma_2 = -.43\}$ $\{\pi_0 = .72, \pi_1 = -.93, \pi_2 = .30\}$

Notes. Results of *k=5* folds and 2000 simulations shown in left panel and *k*=10 folds and 1000 simulations in the right panel.

4.6.3 Discussion

The cross-validation results support the choice of a 1-parameter weighted ΔP rule over the more general 2-parameter weighted ΔP and the 6-parameter linear combination model.

If the 1-parameter model is, in fact, the true population model then we should expect model performance to converge as sample size increases. A large calibration sample would produce parameter estimates for (4.2) and (4.3) that would match the restrictions implied by (4.1), and so MSE for the validation sample would be identical across all models. Further, convergence should occur more quickly for (4.2) since there is only one additional parameter. Yet the above results show that (4.1) and (4.3) are closer in MSE. Why might this be the case?

We know from experiments 1 and 2 that 1-parameter weighted ΔP does not appear to be the correct model for all conditions. In particular, it performs poorly in the generative [8,0,8,0] and the preventive [0,8,0,8] conditions, among others. If the 1-parameter weighted ΔP model is correct for only a majority of the conditions, then the additional flexibility of the linear combination model may allow for better performance as the sample size becomes arbitrarily large. The current sample size may be "intermediate" so that the linear combination model has enough information to surpass the 2-parameter model, but not yet enough to supplant the 1-parameter model.

4.7 General discussion

The results from experiments 1 and 2, as well as the cross-validation study, further bolster confidence in the 1-parameter weighted ΔP model of causal judgment. Furthermore, the above analysis suggests why the weighted ΔP model is empirically stronger than Bayesian models. Namely, *all* weighted ΔP models predict deterministic judgments for generative conditions of the form [0,b,0,c] and [a,0,0,d] and preventive conditions of the form [a,0,c,0] and [0,b,c,0]. In contrast, no reasonable Bayesian model of strength estimation will make such predictions.

Exploratory analysis suggests that medians rather than means better represent typical individual response strategies. Condition means conceal the deterministic bias, thus making judgments appear more consistent with Bayesian predictions. This finding echoes Mozer et al.'s (2008) criticism of Griffiths and Tenenbaum (2006), wherein the aggregate resembles Bayesian inference even though individual judgments do not.

Ideal data would be sufficiently rich so that individual strategies could be modeled instead of group means or medians. Such data would require many repeated measurements per individual, which is not feasible using a sequential presentation format. Datasets with many repeated measurements, such as those in Yeung and Griffiths (2015), use a list presentation format. One possible analysis would be to use cross-validation on the Yeung and Griffiths' data to determine which model best predicts individual judgment strategies. Such an analysis, of course, assumes that similar psychological mechanisms are used across the sequential and list presentation formats. As mentioned in Chapter 1, there are reasons to be skeptical of this assumption (cf. Perales & Shanks, 2008). For instance, list formats seem to be incompatible with any associative updating, as is found in the Rescorla-Wagner model.

Because it is standard to only report condition means, it is difficult to know whether deterministic bias was widespread in previous studies. However, there is suggestive evidence. Buehner and Cheng (1997) reported the highest standard errors as occurring in the generative [8,0,8,0] and preventive [0,8,0,8] conditions. Such findings are consistent with bimodal response distributions that have a large proportion of 0 and 100 strength ratings. In addition, their generative [0,8,0,8] and [8,0,0,8] conditions and preventive [8,0,8,0] and [0,8,8,0] conditions all had very low standard errors. These low standard errors could be due to a high proportion of 0 ratings in the $\widehat{w}_1 = 0$ conditions and 100 ratings in the $\widehat{w}_1 = 1$ conditions.

The empirical priors of Yeung and Griffiths (2015) also appear to be consistent with a large proportion of deterministic ratings. For generative causes, their conditional prior distributions of w_1 look to be bimodal, with a large peak at a strength of 1 and a smaller peak at a strength of 0. Moreover, the bimodality is more pronounced for extreme values of the base rate (w_0 near 0 or 1). This shape aligns with the high deterministic bias conditions from above. Namely, it is necessary for the prior to have this general qualitative shape in order to return posterior predictions that are close to deterministic ratings.

While weighted ΔP predictions were generally affirmed in the analysis, certain conditions did present difficulties. The most problematic were the generative [8,0,8,0] condition and the preventive [0,8,0,8] condition as there appears to be several distinct response strategies. In addition, the deterministic bias also appeared in generative conditions in which the effect always occurred and in preventive conditions in which the effect never occurred. Clearly, the weighted ΔP model, as well as any other point-prediction model, will be unable to describe a mixture of

judgment strategies. The next section sketches one potential solution for this difficulty, though a detailed treatment of this topic will have to wait for another time.

4.7.1 Two-stage inference

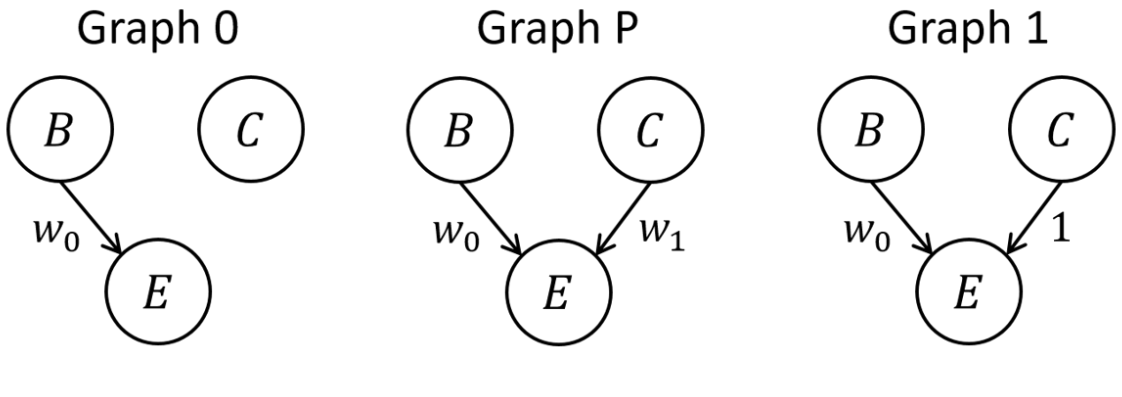
The above discussion referred to causes of two different types. Deterministic causes always produce the effect or never produce the effect while probabilistic causes produce the effect only some proportion of the time. Examples of each type are easy to imagine. When heated to 100 °C water will always boil (assuming standard pressure) and it will never boil when heated to 40 °C. But when eggs are boiled in the water, even when the same procedure is always carefully applied, they will only come out perfect some of the time (perfect, of course, means firm egg whites with a pudding-like and slightly runny yolk). Hence, the preferred egg-boiling procedure generates the effect only with some probability that is less than one.

It is easy enough to conceive of causes as deterministic or probabilistic. Might this distinction also be used in elemental causal induction? One possibility is that causal inference proceeds in two stages. In the first stage, evidence is used to choose between deterministic and probabilistic hypotheses. The second stage then estimates causal strength.

As we saw with the causal support model in Chapter 1, hypotheses may be represented with different graphical models. Consider the three hypotheses corresponding to the three graphs in the top panel of Figure 4.5. Graph 0 and Graph 1 represent deterministic hypotheses for causal strengths of 0 and 1 respectively. And Graph P represents the hypothesis that the cause is probabilistic.

A two-stage inference model can go some of the way towards explaining the deterministic bias and the mixed strategies observed above. The first stage would determine which of the three graphs from Figure 4.5 is most probable given the data. Any inferential procedure could be used, though Bayesian inference is a natural choice. It also potentially gives the normative solution.

Figure 4.5B gives posterior probabilities of the three graphs for the contingency frequencies shown on the horizontal axis. Note that these are the conditions in which the deterministic bias was most often observed. Conditions two through five are ones in which the effect always occurs when the cause is present. For each condition, a large posterior probability is given to Graph 1 and a substantial probability is also given to Graph P. Importantly, the allocation of probability shifts



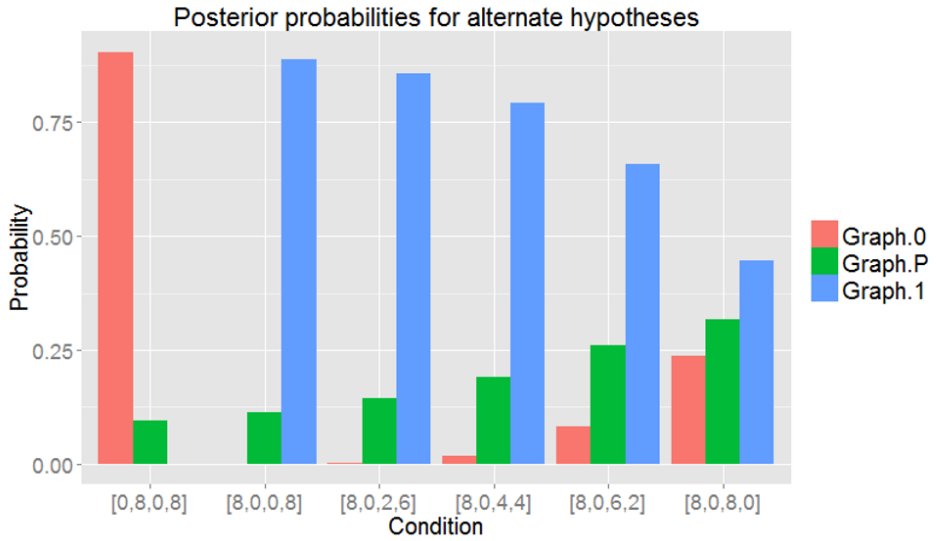


Figure 4.5. Graphs representing competing hypotheses and example posterior probabilities given to each graph for different sets of learning data [a,b,c,d] corresponding to entries from a 2x2 contingency table.

A: Directed graphs with B representing the background variable, C the candidate cause, and E the effect of interest. Graph 0 represents a deterministic hypothesis of no causal strength, or $w_1 = 0$. Graph P represents the hypothesis of probabilistic causal strength with $0 < w_1 < 1$. And Graph 1 represents the deterministic hypothesis of $w_1 = 1$.

B: Posterior probabilities for each graph given the data shown in the condition at the bottom. Condition refers to the [a,b,c,d] cell counts from a 2x2 contingency table. Prior probabilities were set equal to 1/3 for each graph. A Noisy-OR generating function and a joint uniform prior on (w_0, w_1) is assumed for Graph P.

from Graph 1 to Graph P as the observed base rate of the effect increases. Note that this matches the qualitative pattern in the observed distributions of strength judgments (Figure 4.1).

Remember that in the problematic [8,0,8,0] condition, observed strength judgments were about equally divided between 0, probabilistic, and 1. This pattern roughly matches the posterior probabilities given to each hypothesis, shown on the far right of Figure 4.5B. The data in the [8,0,8,0] condition imply that background strength w_0 is large. From Chapter 2 we know that this

further implies high uncertainty for w_1, and the upshot is that all three hypotheses are consistent with the data.

While the two-stage model provides some insight, it alone cannot explain the observed mixture of responses. If all people follow the same two-stage model, then all responses should be homogenous. Additional assumptions are necessary to create a mixture. One possibility is the probability matching hypothesis from Chapter 2, which posits that people take one or a few samples from their posterior instead of selecting the highest probability option. Another possibility is that people have heterogeneous prior beliefs for the different models. Repeated measurements are necessary to discriminate between these two possibilities, as probability matching implies within person variation for a given condition while heterogeneous priors does not.

4.7.2 Bayesian inference after all?

The two-stage model serves another useful purpose in motivating another Bayesian "competitor" model of causal inference. It was asserted above that the data from experiments 1 and 2 are incompatible with all Bayesian models of strength estimation. However, it bears emphasis that the data do not rule out Bayesian models more generally.

A two-stage Bayesian inference model can be constructed to closely mimic the predictions of the weighted ΔP rule. The two stages proceed as before: First, one of three models is selected by choosing the graph with the highest posterior probability. Next, strength is estimated according to the structure of the selected graph. The strength estimation in the second stage may also be made Bayesian. For Graph 0 the posterior expectation will be 0 and it will be 1 for Graph 1. This is because all prior weight is placed on the $w_1 = 0$ hypothesis for Graph 0 and likewise all weight is on the $w_1 = 1$ hypothesis for Graph 1. Graph P requires that some joint prior distribution $f(w_0, w_1)$ be specified. For instance, it could be the joint uniform or SS prior from before.

With a carefully designed prior, the two-stage Bayesian model can be made to resemble the weighted ΔP model. Appendix G describes such a prior, while Figure 4.6 shows the 2-stage Bayesian predictions against the weighted ΔP model. One can see that the predictions are highly similar. With additional effort, it would be possible to make them more similar still. Note that in the key [0,8,0,8] and [8,0,0,8] conditions, both models make the respective predictions of 0 and 1.

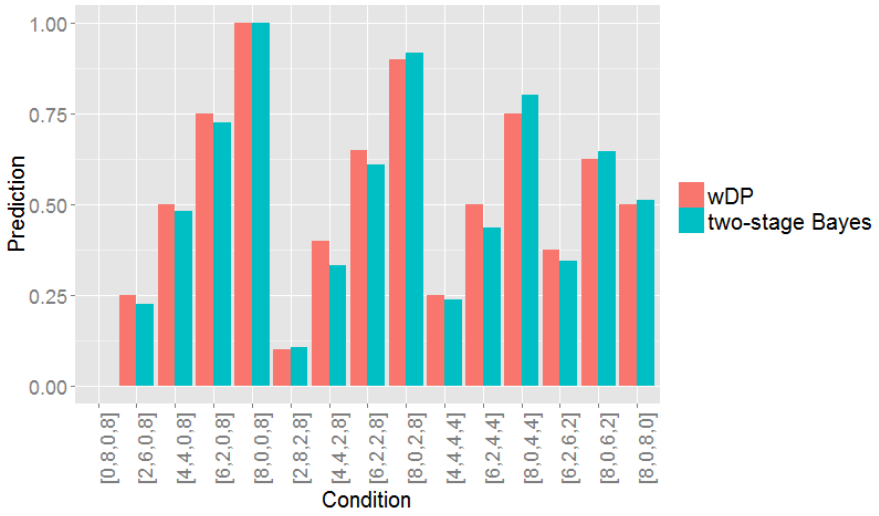


Figure 4.6. Predictions of the weighted Δ*P* and the two-stage Bayesian model for the 15 conditions used in Experiment 1. Note that predictions of 0 and 1 are made respectively for the [0,8,0,8] and [8,0,0,8] conditions.

The two-stage example underscores that the Bayesian approach is a double-edged sword. On one hand, the richness of the framework provides a powerful toolkit for theory building. A potentially valuable hypothesis resulted from construing the causal induction task as a problem of Bayesian model selection. But this example also shows that the extreme flexibility of the approach allows for virtually any data pattern to be fit so long as the researcher is sufficiently creative. This is why critics argue that empirical fit alone is an insufficient criterion for demonstrating the merit of Bayesian models. Additional evidence is necessary.

One form of such evidence, discussed in Chapter 2, is measurements from the environments in which the psychological mechanism is employed. Chapter 5 provides one potential approach for how to characterize the structure of causal environments. It is highly doubtful, however, that environmental measurements will fully resolve the issue. Another form of evidence may come from an elaboration of psychological mechanisms. Can a case be made to prefer certain mechanistic accounts over others? Traditionally, such characterizations have been quite coarse, referring to vague "model simplicity" criteria. In the conclusion I consider recent work that attempts to tighten these criteria so that different mechanisms can be compared in a principled manner.

Chapter 5.

Capacity and response probability: A latent variable framework for models of causal inference

> It is not the laws that are fundamental, but rather the capacities. Nature selects the capacities that different factors shall have and sets bounds on how they can interplay. Whatever associations occur in nature arise as a consequence of the actions of these more fundamental capacities. In a sense, there are no laws of association at all. They are epiphenomena.
>
> (Cartwright, 1989, p. 181).

5.1 Introduction

In *Nature's Capacities and their Measurement*, Nancy Cartwright (1989) argues for the reality of capacities and their fundamental role in causal claims. Capacities describe the power of causes to produce their effects. On Cartwright's account, capacities are fixed and stable from one situation to another. This allows for causal inference to generalize over contexts.

Most natural settings, in Cartwright's view, are characterized by an ever shifting mix of different causes. Strongly uniform or controlled environments, such as those found in the science laboratory, are the exceptions. It is therefore impracticable to learn cause-effect relationships that are specific to a particular environment (she denotes these environment-dependent relationships as *causal laws*). What is needed is causal information that carries from one setting to the next. This is just the role played by Cartwright's capacities.

Cartwright proposes that the concept of capacity is central to standard philosophical accounts of causality, as well as implicit to the statistical methods of many natural and behavioral sciences. One area of her focus is econometrics since the originators of these methods were often explicit in their metaphysical commitments. Cartwright argues that the equations of economics are meant to represent causal relationships, with causes placed on the right side and effects on the left. Causal influence, then, is found in the parameters associated with each cause. Further, in econometrics it

is typical to assume that coefficient values estimated in one context maintain in entirely different contexts. Such practice, Cartwright asserts, evidences a belief in stable capacities that persist across varied settings.

Another fundamental property of Cartwright's capacities is that they are, in general, additive. This property is required if capacities are to do their assigned inferential work. In an important sense, the properties of additivity and stability go hand in hand. To see why, note that causes will be additive if and only if they are independent. Or equivalently, they will be additive if and only if there are no interactions. Now for causal influence to be fixed across contexts, it is essential that there are no interactions. For if there are interactions, the influence of a cause will depend on the levels of the other causes that also occur in a given context. Cartwright emphasizes this point in stating:

> Probably the most common reason for a capacity to fail to obtain in the new situation is causal interaction. The property that carries the capacity interacts with some specific feature of the new situation, and the nature of the capacity is changed. (Cartwright, 1989, p. 163)

Thus, additive relationships are prized since they allow for straightforward generalization of causal influence. Ragnar Frisch and Trygve Haavelmo, two of the founders of econometrics, believed that the parameters would be independent in the fundamental equations of economics (Cartwright, 1989). This position is not surprising since it greatly amplifies the potential utility of econometric analysis. Other prominent economists, including John Maynard Keynes and Robert Lucas, were skeptical of this belief in independent causes. Indeed, both cautioned against econometric predictions due to their concern that parameters would be unstable over varying situations (Cartwright, 1989).

From the preceding discussion, it is easy to see how Cartwright's work serves as an inspiration for Cheng's (1997) power PC model. Recall that power PC provides a context-independent measure of causal strength. Specifically, when a causal power is inferred in one context, that same power will be manifest in novel contexts. Of course, this holds only so long as model assumptions are met. Further recall that a primary assumption of power PC is independence of the candidate and background causes.

Cartwright's paradigmatic examples of capacities at work are linear equations with continuous response. An important contribution of Cheng was to develop a model of independent causation for binary effects. Though independence is central to Cheng's model, additive capacities are not part of her theory. In fact, no previous work in psychology has incorporated additive capacity into models of causal strength judgment.

This chapter integrates additive capacities with judgment models of causal strength. In keeping with Cartwright, I propose it is natural to conceive of causes as having stable capacities that demonstrate fixed influence over heterogeneous contexts. The introduction of a response function is used to link additive capacities to judgments of causal strength. Different response functions are used to represent different judgment models, such as the ΔP rule or causal power.

The proposed framework allows for a broader conception of causal independence. Capacities are fundamental instead of probabilities. If a response function can be found that allows for additive capacities, then causes are independent with respect to that response function. The response function, in turn, determines the probabilities.

After establishing the capacity framework below, I will then sketch several potential applications. Capacities can be used to describe and explore causal intuitions. They might also be employed to measure objective causal environments. These measurements, in turn, could then potentially be used to advance the study of causal learning.

5.2 Capacity as a latent variable

Latent variables are sometimes used to motivate models of dichotomous and categorical outcomes (e.g. Long, 1997). While a dichotomous variable is observed to take only one of two values, it is often reasonable to consider such outcomes as some function of a continuous underlying propensity. For example, we may observe that two people die of cardiac arrest. But for this same observed outcome, the two people could differ with respect to their underlying risk. One could be an otherwise healthy person while the other could be a heavy smoker.

In the proposed framework, causes have unobserved capacities that determine the probability of the effect. Capacity is a scalable, latent variable that is mapped to probability by a response function. I will often refer to the latent variable framework as CARP, which is an acronym for *capacity and response probability*. The general details of CARP are described in the next section.

5.2.1 Definitions and assumptions

Assume a set of dichotomous causes $\{C_i: i \in I\}$, a context or background cause B, and a dichotomous effect E. Notation is consistent with before wherein c_i^+, b^+, and e^+ indicate the occurrence of cause C_i, background B, and effect E while c_i^-, b^-, and e^- represent their absence. In keeping with the previous account, the background B is unobserved, but assumed to be always present. For a given set $\{C_i: i \in I\}$ causes are distinguished as either *single* or *conjoined*. A single cause C_i is just one member from the set while a conjoined cause $C_{i,j} = (C_i \,\&\, C_j)$ denotes two or more single causes.

Under the proposed framework, causes are associated with *capacities*. In particular, a given cause C_i is associated with a capacity α_i. Conjoined causes are assumed to combine additively in their capacities so that $(C_i \,\&\, C_j) \rightarrow \alpha_i + \alpha_j$. The background cause B also has an additive contribution to capacity, which will typically be denoted with α_0. So the total causal capacity of cause C_i in the generic context B is $(B \,\&\, C_i) \rightarrow \alpha_0 + \alpha_i$.

The link between capacities and probabilities is achieved through a *response function*. A response function $F(.)$ applied to a capacity α determines a probability $w = F(\alpha)$. Response functions are assumed to be strictly increasing and continuous in capacity. As such, each response function will have an inverse. The inverse response function $F^{-1}(.)$ returns a capacity when applied to a probability so that $\alpha = F^{-1}(w)$.

The *causal strength* of cause C_i is defined as the probability $w_i = F(\alpha_i)$. So the strength of a given cause C_i is the hypothetical probability that results from adding its capacity α_i to a context with a capacity of 0. This is simply because $F(\alpha_i) = F(\alpha_i + 0)$. For strengths to be informative, then, it will require that $F(.)$ be increasing in the neighborhood of zero capacity.

The above definition of causal strength is compatible with the graphical definition from Chapter 1, wherein the causal strength of cause C_i corresponds to its edge weight w_i in a common effect causal graph (Figure 5.1).

It will become evident that it is the capacities that are the true carries of causal information, not the causal strengths. The definition of causal strength with respect to a zero capacity reference context is, in some sense, arbitrary. However, defining strength in this way facilitates the connection to psychological models, as will be seen shortly.

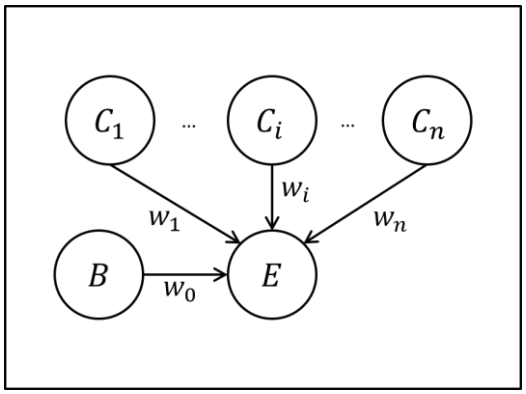


Figure 5.1. A common effect causal graph with a binary effect *E*, a background cause *B* with causal strength w_0, and an arbitrary number of candidate causes C_i with causal strengths w_i. The background *B* is assumed to be always present.

Cumulative distribution functions (cdfs) can serve as response functions, so long as they are continuous and strictly increasing. The framework could be generalized to cdfs with points of discontinuity, though there would be complications in determining the inverse. Non-densities may also be used for response functions, though predicted judgments will no longer conform to the rules of probability.

A response function and its inverse are all that are required for the framework. However, I will often assume that the response function has a derivative $f(\alpha) = \frac{d}{d\alpha}F(\alpha)$ and that the derivative is continuous. The derivative of a response function $F(\alpha)$ is referred to as the *response curve* $f(\alpha)$. When a cumulative density is used for the response function, the response curve is given by the corresponding probability density function.

To summarize, response functions form the link between additive capacities and effect probabilities. Different functions can be constructed to represent different beliefs about the causal system. Response functions can either be specified a priori or estimated from data. In the sections to follow, I will find the a priori response functions implied by the ΔP rule and the power PC model.

Alternatively, one might attempt to find response functions from empirical measurements. Subjective response functions could be estimated from human judgments, though it would require many observations with low measurement error. One could also estimate objective response functions using measurements from actual causal environments. Both of these possibilities are explored later in the chapter.

5.2.2 Elemental causal induction

I will now demonstrate how CARP may be applied to the problem of elemental causal induction. This section describes a general procedure for obtaining strength judgments from response functions. Once the procedure is established, I will use it in Section 5.3 to find the response functions implied by the ΔP rule and the power PC model.

As before, assume that the effect is caused either by an unobserved background cause B or a candidate cause C (as in Figure 1.1). For now, also assume that probabilities are observed without error. Then the probability $w_0 = P(e^+|c^-, b^+)$ gives the causal strength of the background B while $w_T = P(e^+|c^+, b^+)$ is the conjoined strength of B and C.

For a given response function $F(.)$ the causal strength of C can be found using the three steps below:

Step 1: Infer the capacities for B and $(B\&C)$ with $\alpha_0 = F^{-1}(w_0)$ and $\alpha_T = F^{-1}(w_T)$.

Step 2: Find the capacity of the candidate cause C with $\alpha_1 = \alpha_T - \alpha_0$.

Step 3: Find the causal strength of C with $w_1 = F(\alpha_1)$.

Steps 1 follows directly from the definition of a response function. Step 2 follows from the additivity assumption, which asserts that $\alpha_T = \alpha_0 + \alpha_1$. And step 3 is just the definition of causal strength from above.

The mechanics of CARP can be illustrated with a graphical example. Figure 5.2 shows an example response curve, given by a beta(1,3) density function. The response curve shows that initial increments in capacity produce large increases in probability, though as capacity accumulates there are diminishing gains.

For this example, assume a background rate of $w_0 = 0.75$ while the probability with cause present is $w_T = 0.95$. The 3-step procedure for finding causal strength is shown in Figure 5.3. The causal capacities for the background cause B and the candidate cause C are found with the response

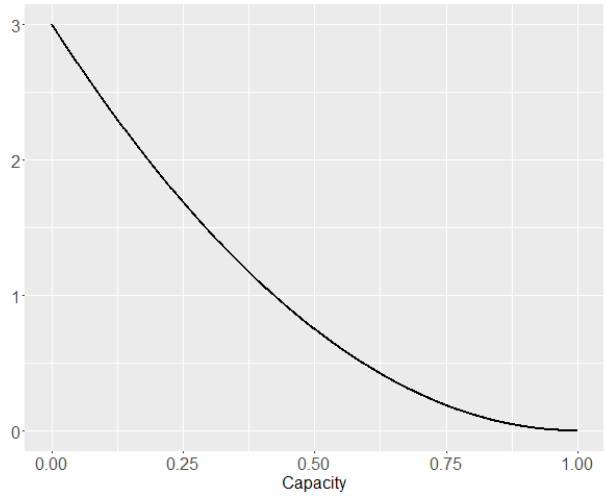


Figure 5.2. Example of a response curve given by a beta(1,3) density function. Latent causal capacity is measured on the abscissa. For a given capacity the probability of the effect corresponds to area under the curve.

curve from Figure 5.2. For a beta(1,3) density, the inferred capacities are $\alpha_0 = F^{-1}(0.75) = 0.37$ and $\alpha_T = F^{-1}(0.95) = 0.63$. This implies that $\alpha_1 = \alpha_T - \alpha_0 = 0.26$. The strength of the candidate cause is then $w_1 = F(\alpha_1) = 0.6$.

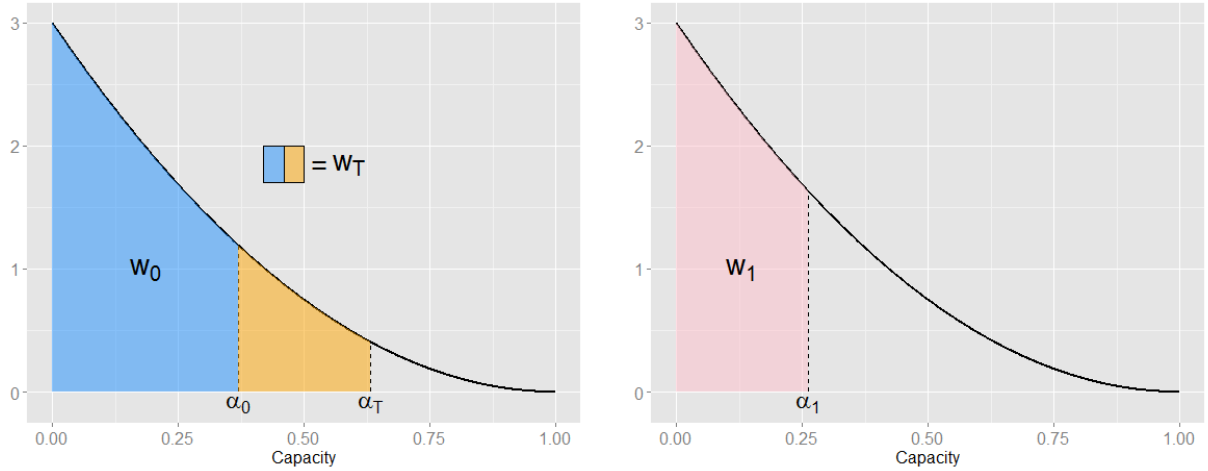


Figure 5.3. Causal capacities and causal strengths inferred from assuming a beta(1,3) response curve.

A: The causal strength w_0 for the background cause B is given by the area in blue. The causal strength w_T for the combined cause $(B\&C)$ is given by the area in blue plus the area in orange. The observed probabilities can be used to infer the capacity α_0 for B and the combined capacity α_T for the conjunction of causes $(B\&C)$.

B: The causal capacity of the candidate cause is found with $\alpha_1 = \alpha_T - \alpha_0$. Causal capacity can then be used to find the causal strength of C, which is given by the pink area.

Step 3 of the procedure is not essential: once causal capacity is determined, the causal influence of C can be found with respect to any context. Again, step 3 is useful for developing the connection between the latent variable framework and psychological models, a task pursued in the next section.

5.3 CARP representation of rational models

This section will demonstrate how the leading rational models of causal inference can be represented using the latent variable framework. The general strategy is to begin with a particular model's definition of causal strength, and substitute it into the 3-step procedure from the previous section. This allows one to find the response curve implied by that model.

5.3.1 The ΔP rule

Begin with the definition of causal strength given by the ΔP rule. Expressed in weights notation it is:

$$\mathrm{w}_1 = w_T - w_0$$

From the 3-step procedure, causal strength can generally be expressed as:

$$\begin{aligned} \mathrm{w}_1 &= F[F^{-1}(w_T) - F^{-1}(w_0)] \\ F^{-1}(\mathrm{w}_1) &= F^{-1}(w_T) - F^{-1}(w_0) \end{aligned} \tag{5.1}$$

Now plug in the ΔP definition of w_1 to obtain:

$$F^{-1}(w_T - w_0) = F^{-1}(w_T) - F^{-1}(w_0) \tag{5.2}$$

It is evident that any $F^{-1}(w) = kw$ for $k \neq 0$ will satisfy (5.2). Choosing $k = 1$ implies $F(\alpha) = \alpha$ and $\frac{d}{d\alpha}F(\alpha) = f(\alpha) = 1$. If one also assumes a proper density and a $[0,1]$ interval of support, then $f(\alpha)$ is a beta(1,1) or uniform density. The response curve, the response function, and the inverse response function are then:

$$f(\alpha) = \begin{cases} 1 & \text{for } 0 \leq \alpha \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$F(\alpha) = \begin{cases} 0 & \text{for } \alpha < 0 \\ \alpha & \text{for } 0 \leq \alpha \leq 1 \\ 1 & \text{for } \alpha > 1 \end{cases} \qquad F^{-1}(w) = w \text{ for } 0 \leq w \leq 1$$

Figure 5.4 shows the response curve for the ΔP model with $w_0 = 0.3$ and $w_T = 0.5$. It bears emphasis that it is the shape of the response curve that determines the model while the scale of causal capacity is arbitrary. A uniform density over any interval $[a, b]$ with $k = \frac{1}{b-a}$ could be chosen. However, for step 3 from above to be sensible, it should be modified so that causal strength is given by $w_i = F(a + \alpha_i)$ for a given interval $[a, b]$.

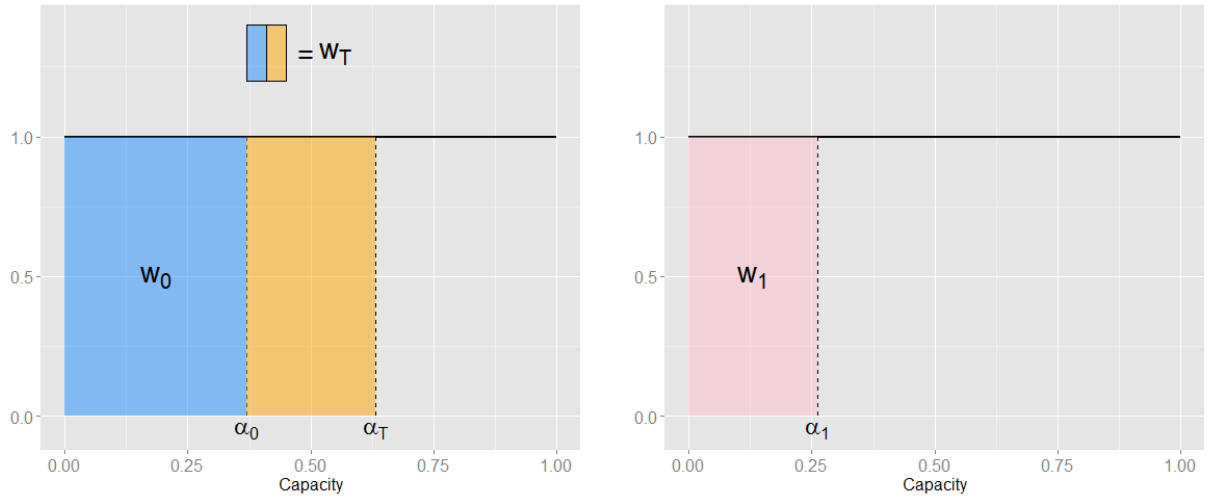


Figure 5.4. A uniform density response curve, which gives causal strengths of the ΔP rule.

A: The background causal strength $w_0 = 0.3$ is given in blue while the conjoined strength $w_T = 0.5$ is the area in blue plus the area in orange.

B: The causal capacity of the candidate cause is found with $\alpha_1 = \alpha_T - \alpha_0 = 0.5 - 0.3 = 0.2$. The capacity is then used to find the strength of the candidate cause $F(\alpha_1) = w_1 = 0.2$, which is the area in pink.

5.3.2 Causal power

We now seek a response function for the causal power model. Recall the expression for generative causal power:

$$w_1 = \frac{w_T - w_0}{1 - w_0}$$

Plugging causal power into (5.1) gives:

$$F^{-1}\left(\frac{w_T - w_0}{1 - w_0}\right) = F^{-1}(w_T) - F^{-1}(w_0) \tag{5.3}$$

Hence, we need to find the function $F^{-1}(.)$ that satisfies (5.3). To find this function, define $F^{-1}(w) = -G(1 - w)$. Then the left-hand side of (5.3) can be expressed as:

$$\begin{aligned} F^{-1}\left(\frac{w_T - w_0}{1 - w_0}\right) &= -G\left(1 - \frac{w_T - w_0}{1 - w_0}\right) \\ &= -G\left(\frac{1 - w_T}{1 - w_0}\right) \end{aligned}$$

And the right hand side terms of (5.3) are:

$$F^{-1}(w_T) = -G(1 - w_T) \qquad\qquad F^{-1}(w_0) = -G(1 - w_0)$$

Putting it all together:

$$\begin{aligned} -G\left(\frac{1 - w_T}{1 - w_0}\right) &= -G(1 - w_T) + G(1 - w_0) \\ G\left(\frac{1 - w_T}{1 - w_0}\right) &= G(1 - w_T) - G(1 - w_0) \end{aligned} \tag{5.4}$$

It is evident that $G(w)$ must be a logarithmic function since $\log\left(\frac{A}{B}\right) = \log(A) - \log(B)$. From the definition of $G(w)$ this implies that $F^{-1}(w) = -\log(1 - w)$. Also, since $w \in [0,1]$ it implies that capacity is defined only over non-negative numbers, or $\alpha \in [0, \infty)$. Any base for the logarithm may be used to satisfy (5.3). The natural logarithm is a convenient choice since it allows for a clean expression of the probability density function.

Using the natural log, we can derive the response curve, the response function, and the inverse response function:

$$f(\alpha) = \begin{cases} e^{-\alpha} & \text{for } \alpha \geq 0 \\ 0 & \text{for } \alpha < 0 \end{cases}$$

$$F(\alpha) = \begin{cases} 1 - e^{-\alpha} & \text{for } \alpha \geq 0 \\ 0 & \text{for } \alpha < 0 \end{cases} \qquad F^{-1}(w) = -\log(1 - \mathrm{w}) \qquad 0 \leq \mathrm{w} < 1$$

Thus, an exponential density with $\lambda = 1$ will give a latent variable model for causal power. Again, the scale is arbitrary, so choosing any exponential density (i.e. any $\lambda > 0$) will also give the causal power model, which is shown in Appendix H. The choice of the logarithm base in the above derivation is equivalent to choosing λ, with base e corresponding to $\lambda = 1$.

Figure 5.5 shows an exponential response curve and inference for $w_0 = 0.75$ and $w_T = 0.95$. Hence, CARP provides an alternate conception of causal power. Causal power is inferred causal strength with respect to an exponential response curve.

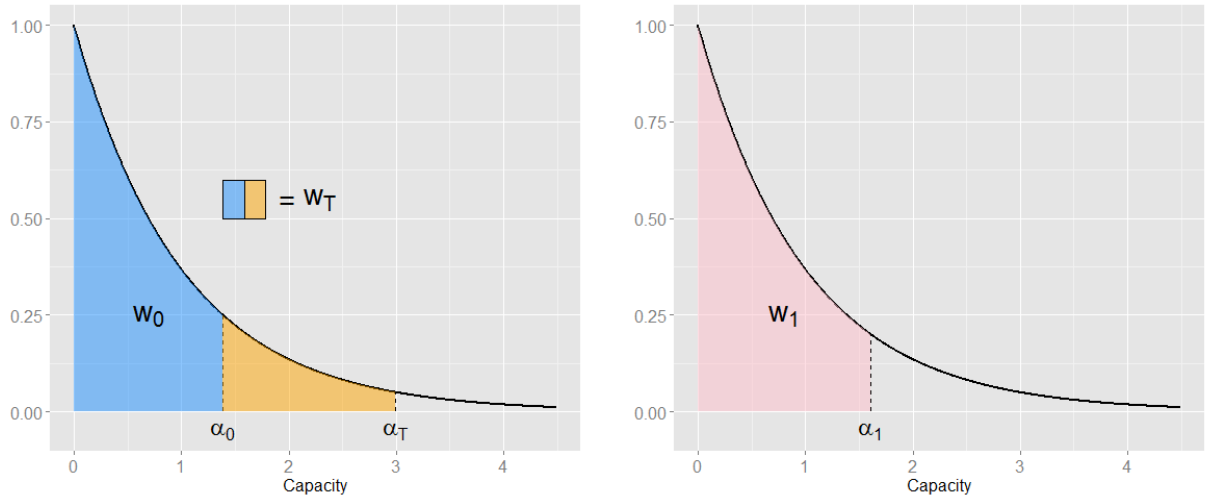


Figure 5.5. An exponential density response curve, which gives causal strengths of the causal power model.

A: The background causal strength $w_0 = 0.75$ is given in blue while the conjoined strength $w_T = 0.95$ is the area in blue plus the area in orange.

B: The causal capacity of the candidate cause is found with $\alpha_1 = \alpha_T - \alpha_0 \approx 3 - 1.4 = 1.6$. The capacity is then used to find the strength of the candidate cause $F(\alpha_1) = w_1 = 0.8$, which is the area in pink.

5.3.3 Relationship between ΔP and causal power

A superficial comparison shows that the exponential curve of Figure 5.5 is similar to the beta(1,3) curve shown in Figure 5.3. The inferred strengths from the two models are also similar, with the beta(1,3) model giving a strength of about 0.6 while the exponential model returns a strength of

0.8. This similarity points towards a deeper relationship that connects the ΔP rule to the causal power model.

A uniform or beta(1,1) density gives causal strengths of the ΔP model, an exponential density produces causal power, and the beta(1,3) density returns predictions intermediate of the two models. It can be shown that $\lim_{n\to\infty} \text{beta}(1,n) \to \exp(\lambda = 1)$ (see Appendix H). Thus, the ΔP rule and causal power are particular cases that lie along a continuum of latent variable models.

One potential advantage of the $\text{beta}(1,n)$ formulation of causal power is that it allows finite capacities to give deterministic strengths. Namely, any capacity $\alpha_i > 1$ will give $w_i = 1$. In contrast, the exponential formulation allows for deterministic causal strengths of 1 only when causal capacity is infinite.

The beta density is quite flexible and may also be used to approximate other proposed models of causal judgment. The "Per Cent Success" rule, covered in Chapter 1, posits that people completely neglect the base rate to give a causal judgment of $w_1 = w_T$. This model can also be approximated with a beta density response curve by using $\lim_{\epsilon\to 0} \text{beta}(\epsilon, 1)$.

5.3.4 Preventive causes

The latent variable framework may be extended to accommodate preventive causes. Preventive causes are defined as those that reduce the probability of the effect. Suppose context B has a capacity of α_0. Then cause C_i will be preventive in context B if $F(\alpha_0 + \alpha_i) \leq F(\alpha_0)$. From the definition of $F(.)$, this implies $\alpha_i \leq 0$ for preventive causes.

Within the causal power paradigm, preventive strength is defined with the following hypothetical: Suppose there is a context in which the effect always occurs. When preventive cause C_i is introduced to that context, the probability of the effect is reduced to $1 - w_i$, and the preventive strength of C_i is w_i.

The CARP framework will follow this causal power convention for preventive strength. Namely, preventive strength is defined as a positive quantity subtracted from a reference probability of 1. It then becomes necessary to amend the above 3-step procedure to match this conception. When C is a preventive cause such that $w_T < w_0$, step 3 becomes:

> **Step 3 (preventive):** Choose b so that $F(b) = 1$. Then preventive strength of C is given by $w_1 = 1 - F(b + \alpha_1)$.

When $F(.)$ is a continuous cumulative density function, as is assumed below, then b will be the minimum number such that $F(b) = 1$. This will ensure that $F^{-1}(1)$ is unique and equal to b. The amendment limits the class of available response functions to those with support bounded above. Otherwise, $b = \infty$ for $F(b) = 1$ and all preventive strengths would be zero.

5.3.4.1 Preventive ΔP

The same general strategy can be used to find the preventive ΔP response function, but now using the amended step 3. First, a small revision must be made to the ΔP rule to achieve consistency with CARP. For $w_T < w_0$ the ΔP rule gives negative causal strengths. To obtain agreement with CARP's positive preventive strengths, the ΔP rule can be defined as:

$$\mathrm{w}_1 = \begin{cases} w_T - w_0 & \text{for } \mathrm{w_T} \geq w_0 \\ -(\mathrm{w_T} - w_0) & \text{for } \mathrm{w_T} < w_0 \end{cases}$$

Now suppose that for the response function $F(.)$ we can find a b so that $F(b) = 1$. Then preventive causal strength for cause C_1 is:

$$\begin{aligned} w_1 &= 1 - F(b + \alpha_1) \\ &= F(b) - F(b + \alpha_1) \end{aligned}$$

Recall from step 2 above that $\alpha_1 = F^{-1}(w_T) - F^{-1}(w_0)$. Also, for preventive ΔP we have causal strength defined as $\mathrm{w}_1 = -(\mathrm{w_T} - w_0)$. Putting this all together:

$$\begin{aligned} -(\mathrm{w_T} - w_0) &= 1 - F[b + F^{-1}(w_T) - F^{-1}(w_0)] \\ F^{-1}[1 + (w_T - w_0)] &= b + F^{-1}(w_T) - F^{-1}(w_0) \\ F^{-1}[1 + w_T - w_0] &= F^{-1}(1) + F^{-1}(w_T) - F^{-1}(w_0) \end{aligned}$$

So once again, $F(.) = kx$ for $k > 0$ will serve as the class of response functions. Thus, any cumulative uniform density will do the job. If we choose the density on the $[0,1]$ interval, then $b = 1$ satisfies the requirement of the minimum b with $F(b) = 1$. Figure 5.6 shows the preventive ΔP latent variable model.

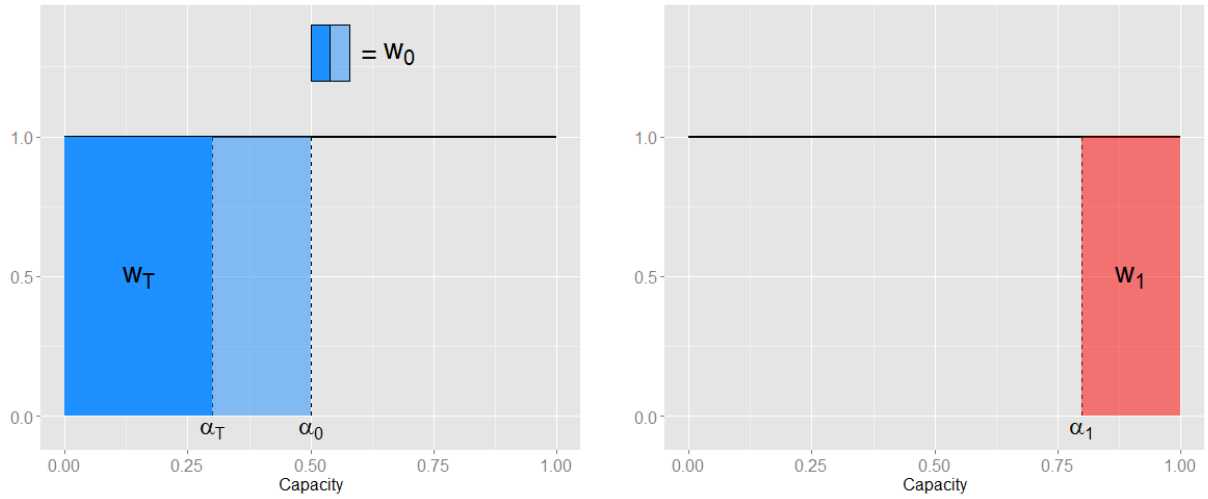


Figure 5.6. A uniform density response curve, which gives causal strengths of the preventive ΔP rule.

A: The background causal strength $w_0 = 0.5$ is given by the dark blue area plus the light blue area. The conjoined strength $w_T = 0.3$ is given by the dark blue area. Since $w_T < w_0$, the cause is preventive.

B: The causal capacity of the candidate cause is found with $\alpha_1 = \alpha_T - \alpha_0 = 0.3 - 0.5 = -0.2$. The capacity is then used to find the strength of the candidate cause with $1 - F(1 + \alpha_1) = w_1 = 0.2$, which is the area in red.

5.3.4.2 Preventive causal power

Now we find the response function for preventive power. Assume that there is some b such that $F(b) = 1$. Then for preventive power:

$$\frac{-(w_T - w_0)}{w_0} = 1 - F[b + F^{-1}(w_T) - F^{-1}(w_0)]$$

$$F^{-1}\left(\frac{w_T}{w_0}\right) = b + F^{-1}(w_T) - F^{-1}(w_0)$$

$$= F^{-1}(1) + F^{-1}(w_T) - F^{-1}(w_0)$$

From the final line it is apparent that $F^{-1}(.)$ must be a logarithm. As before, choose the natural log with base e. The response curve, the response function and the inverse response function are then:

$$f(\alpha) = \begin{cases} e^{\alpha} & \text{for } \alpha \leq 0 \\ 0 & \text{for } \alpha > 0 \end{cases}$$

$$F(\alpha) = \begin{cases} e^{\alpha} & \text{for } \alpha \leq 0 \\ 0 & \text{for } \alpha > 0 \end{cases} \qquad F^{-1}(w) = \log(w) \qquad 0 \leq w < 1$$

The response function has positive support on $[-\infty, 0]$. In addition, $b = 0$ satisfies the requirement of the minimum b with $F(b) = 1$. So for a preventive cause C_1, preventive power is found with $w_1 = 1 - e^{\alpha_1}$.

Figure 5.7 shows the causal response function and inference for the preventive causal power model. The preventive power response curve is the reflection of the generative curve over the $y = 0$ axis. Thus, generative and preventive power require two different response curves. For ΔP, in contrast, both generative and preventive causes may be represented with a single response curve.

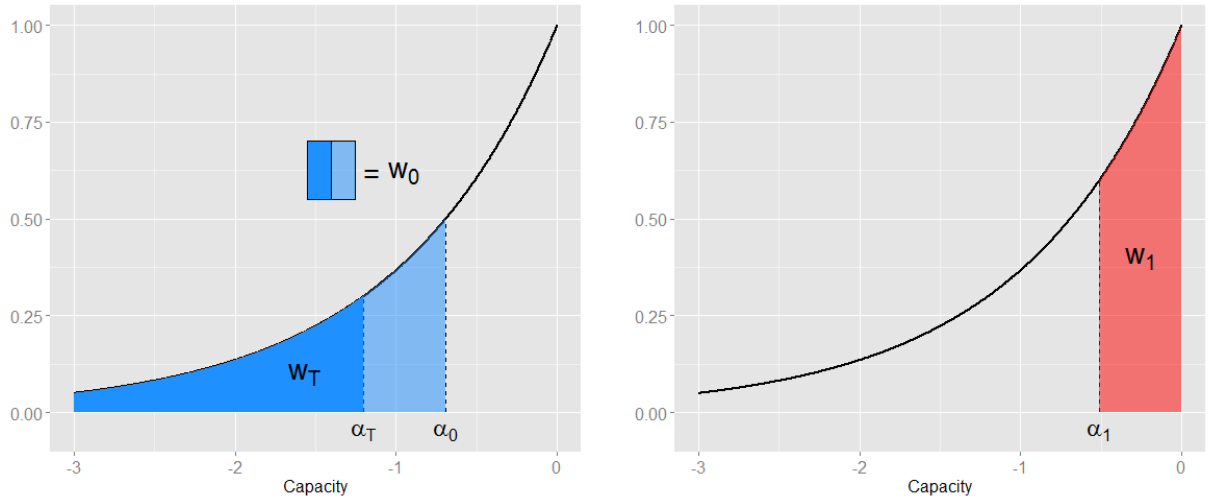


Figure 5.7. An exponential density response curve, which gives causal strengths for preventive causal power

A: The background causal strength $w_0 = 0.5$ is given by the dark blue area plus the light blue area. The conjoined strength $w_T = 0.3$ is given by the dark blue area. Since $w_T < w_0$, the cause is preventive.

B: The causal capacity of the candidate cause is found with $\alpha_1 = \alpha_T - \alpha_0 = 0.3 - 0.5 = -0.2$. The capacity is then used to find the strength of the candidate cause with $1 - F(\alpha_1) = w_1 = 0.2$, which is the area in red.

5.3.4.3 Beta approximation for preventive power

As with the generative model, a beta distribution may be used to approximate the response curve for causal power. Specifically, it can be shown that $\lim_{n\to\infty} \text{beta}(n, 1)$ will approach e^{α}. One can set $b = 1$ to satisfy $F(b) = 1$. Preventive power for a cause C_1 can then be approximated with the expression $w_1 = 1 - F^{-1}(1 + \alpha_1)$.

In summary, CARP shows how additive capacities map to evaluations of causal strength. As such, it provides a new perspective by which to explore models of causal learning. The next sections sketch a number of applications facilitated by the latent variable formalism.

5.4 Conjunctive causation

Novick and Cheng (2004) extend causal power theory to situations in which two candidate causes generate or prevent the effect. They argue that previous theories of interactions are purely covariational because they rely on differences of observed probabilities only. Novick and Cheng, in contrast, embed the concept of an interaction within causal power theory.

The basic idea of Novick and Cheng's approach is to find the causal powers of the two individual causes first, which they refer to as the "simple" powers of the two causes. The interaction is then treated as a separate causal entity for which "conjunctive" power may be found. Novick and Cheng provide a number of formulas to cover the various possible cases (e.g. the case with two generative causes and a preventive interaction).

CARP offers an alternative approach for modeling the influence of multiple causes. As mentioned above, the additive capacity assumption is equivalent to an assumption of no interactions between causes. Such an assumption might seem overly restrictive. However, some flexibility is recovered by allowing the response curve to be adjusted to the data. Furthermore, the additivity assumption allows for a more parsimonious representation of a system with two or more causes. It also permits easy generalization to causes measured on a continuous scale.

The contrast between approaches is best illustrated with an example. Consider the factors or "causes" that lead to high incomes. Occupation is a clearly important, as some job types have a much higher proportion of high incomes than others. The sex of the worker is also important, as men have typically been paid more than women.

Figure 5.8 shows the common effect graph for this example. The example uses income data

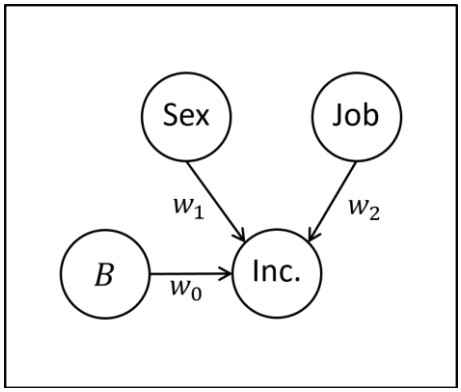


Figure 5.8. Common effect causal graph for causes of high income. Cause 1 is sex ($c_1^+ =$ male), cause 2 is job type ($c_2^+ =$ professional or white collar job) and effect is "makes over $50k per year".

from the 1994 Census. The effect will be defined as "makes over $50,000 a year". Both causes are coded to be generative with $c_1^+ =$ male and with $c_2^+ =$ professional or white collar job.

Now the causal power model will hold if all the causes are independent. But what does this mean exactly? An important distinction can be made between independence of *occurrence* versus independence of *influence*. With observational data, two causes may be dependent because they tend to co-occur. That is, the causes may be statistically dependent. This is less of an issue with experimental data as a balanced design will eliminate dependence of observed factors and random assignment will eliminate dependence with unobserved variables. Causes can also be dependent in their influence on an outcome. This second form of dependence corresponds to the notion of an interaction. To use one of Novick and Cheng's (2004) examples, hard work alone and talent alone are weak causes of success. However, the combination of these causes generates a high probability of success. So this is an example of a positive interaction.

The causal power model assumes that the various causes of an effect are independent in both occurrence and influence (Appendix A). Figure 5.8 represents the assumption of statistical independence since there are no edges connecting the causal variables. In fact, for the Census example the two candidate causes are essentially statistically independent with $\Pr(c_2^+|\text{male}) = 0.26 \approx \Pr(c_2^+|\text{female}) = 0.27$. This should reassure us that the Figure 5.8 is not too great a distortion of reality. Clearly, the candidate causes are not randomly assigned and so there may be dependence with background factors, though we ignore this complication for now.

Next consider the interaction. To streamline the example, assume that the observed probabilities are population quantities. The probabilities are:

$$w_0 = \Pr\left(e^+ \middle| c_{sex}^-, c_{job}^-\right) = .06 \qquad w_{T1} = \Pr\left(e^+ \middle| c_{sex}^+, c_{job}^-\right) = .22$$

$$w_{T2} = \Pr\left(e^+ \middle| c_{sex}^-, c_{job}^+\right) = .25 \qquad w_{T12} = \Pr\left(e^+ \middle| c_{sex}^+, c_{job}^+\right) = .57$$

With regards to notation, w_0, w_1 and w_2 represent the simple powers of the background cause, cause 1, and cause 2, respectively. The w_{T1} gives the total probability resulting from the combined influence of the background and cause 1 while w_{T2} has a symmetric definition for cause 2. The w_{T12} gives the total probability from the background, cause 1, and cause 2. The "conjunctive power" of cause 1 and cause 2 is represented by w_{12}, which describes the interaction of the two causes.

5.4.1 Interactive causal power

Begin with the approach of Novick and Cheng (2004). The first step is to find the simple powers of sex and job type with:

$$w_1 = \frac{(w_{T1} - w_0)}{1 - w_0} = .17$$

$$w_2 = \frac{(w_{T2} - w_0)}{1 - w_0} = .20$$

So being male and obtaining a professional job are both generative causes of a high income. Next is to find the expected probability of the conjunction if the causes exert only their simple powers. This expectation forms the "no interaction" baseline within the causal power model. It is given by:

$$E[w_{T12}] = 1 - (1 - w_1)\times(1 - w_2)\times(1 - w_0)$$

One can see that the formula gives the complement to the event that cause 1, cause 2, and the background cause all do not generate the effect, assuming each of these events are independent. Using the above probabilities yields $E[w_{T12}] = .38$. The next step is to compare the expected probability assuming simple powers only to the probability actually observed. In this example, the observed conjunctive probability is $w_{T12} = .57$ so $E[w_{T12}] < w_{T12}$. Hence, conjunctive power is generative and given by the equation:

$$w_{12} = \frac{(w_{T12} - E[w_{T12}])}{1 - E[w_{T12}]}$$

This example gives a conjunctive power of $w_{12} = .32$, representing a positive interaction between sex and job type for determining income.

5.4.2 CARP formulation of interactive power

Even with just two causes, the formulas for interactive causal influence become rather complex. This section shows how Novick and Cheng's model can be expressed within CARP. One advantage of CARP is that it allows for a somewhat more compact representation.

Let $G(.)$ represent the response function for generative power. Then causal capacities for the background, sex, and job type can be found as before:

$$\alpha_0 = G^{-1}(w_0)$$
$$\alpha_1 = G^{-1}(w_{T1}) - G^{-1}(w_0)$$
$$\alpha_2 = G^{-1}(w_{T2}) - G^{-1}(w_0)$$

The assumption of additive capacities is equivalent to the assumption of no interaction, or "simple powers only" under the causal power model. Within CARP, the interaction may be assessed according to capacities instead of probabilities. The expected conjunctive capacity is $E[\alpha_{T12}] = \alpha_0 + \alpha_1 + \alpha_2$ while the observed conjunctive capacity is:

$$\alpha_{T12} = G^{-1}(w_{T12})$$

An interaction occurs if the observed conjunctive capacity does not equal expected capacity. If $\alpha_{T12} > E[\alpha_{T12}]$, as in the example, then conjunctive power is generative. The generative power of the interaction is then found with:

$$w_{12} = G[\alpha_{T12} - (\alpha_0 + \alpha_1 + \alpha_2)]$$

And the obtained w_{12} will be the same as the one found with Novick and Cheng's model. The CARP representation is straightforward in this case because all of the causes, as well as the interaction, are generative.

If observed capacity had been less than expected, with $\alpha_{T12} < E[\alpha_{T12}]$, then conjunctive power would have been preventive. This case becomes more complicated because it requires shifting to the preventive response function. As a result, it is no longer possible to work with capacities only. Instead, one must find the expected conjunctive probability with:

$$E[w_{T12}] = G[\alpha_0 + \alpha_1 + \alpha_2]$$

Then preventive interactive power is found by treating the expected probability as the base rate while using a preventive response function. Specifically, let $R(.)$ be the preventive response function for causal power. Then preventive interactive power is:

$$w_{12} = R[R^{-1}(w_{T12}) - R^{-1}(E[w_{T12}])]$$

Proceeding in this manner, one can find the CARP representation of all 6 cases described by Novick and Cheng. In general, as generative and preventive causes are mixed, a causal power analysis becomes more difficult. This is because one must shift between the generative and preventive response curves from above.

5.4.3 Causal systems perspective

The last section showed that CARP may be used to model interactions within the causal power framework. However, the procedure is fairly involved using either Novik and Cheng's (2004) algebraic expressions or CARP's response function formalism. Surely, finding an interaction under any model will be complex as it requires: 1) the simple effects of the two candidate causes, 2) the expected probability from the simple effects alone and 3) the interaction, which is some function of the observed and expected probabilities. Such an involved procedure seems an unlikely candidate for a psychological process. Might there be a different representation that simplifies the process while still tracking the causal strengths?

As Novick and Cheng (2004) observe, an interaction is a model dependent term. Accordingly, the same evidence can imply an interaction with respect to one causal model and no interaction with respect to a different model. So one potential strategy for simplifying inference is to adopt a model that minimizes or eliminates interactions for certain domains of evidence. This may be done within CARP since different response curves correspond to different models of the causal system.

To see how this can work, return to the Census example from above. The response curve in Figure 5.9 is constructed so that additivity, and therefore independence, is satisfied for the background cause and the observed causes of sex and job type. This means that the capacity of the conjunction is the sum of the simple capacities. Then for the Census example, $F(\alpha_0 + \alpha_1 + \alpha_2) = F(\alpha_{T12}) = .57$, where $F(.)$ is a beta(1, .34) density. Many other response functions could have been chosen so long as they meet the additivity constraint. This particular response function was found by setting the first shape parameter equal to 1 and then searching for the second parameter that satisfied additivity.

The response function must have at least one free parameter to fit the Census data, though this

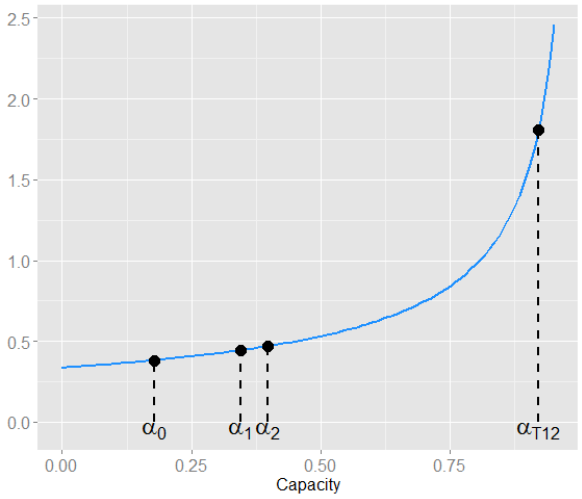


Figure 5.9. A beta(1 , .34) density response curve. The background is all workers who are female and not white collar or professionals. Cause 1 is sex with c_1^+ = male and cause 2 is job type with c_2^+ = white collar or professional. The response curve was constructed to satisfy additivity of capacities so that $\alpha_{T12} = \alpha_0 + \alpha_1 + \alpha_2$.

is not the only requirement. It must also be convex on at least part of the interval between α_0 and α_{T12}. The reason is because $F(\alpha_0) + F(\alpha_{T1}) + F(\alpha_{T2}) < F(\alpha_{T12})$, which by additive capacities implies that:

$$F(\alpha_0) + F(\alpha_1) + F(\alpha_2) < F(\alpha_0 + \alpha_1 + \alpha_2)$$

And so $F(.)$ must be accelerating, or convex, on at least part of the $[\alpha_0, \alpha_{T12}]$ interval. Equivalently, the response curve $f(.)$ must be increasing on part of the interval. Clearly, this is satisfied by the curve shown in Figure 5.9 since it is increasing over its entire support.

Hence, an appropriately chosen causal model can eliminate interactions and simplify inference for conjunctive causation. Of course, it is far from parsimonious to specially tailor a model just to eliminate an interaction for one set of probabilities. However, complete elimination of the interaction may not be necessary for good prediction. Instead, one could adopt just a few paradigmatic causal models for different types of evidence. Causal power may serve as a good default model in most situations, though reasoners may select another standard model for other types of contexts.

The type of model shown in Figure 5.9 describes a causal system governed by "weak joint necessity" wherein each cause alone has small or negligible impact while the combination of causes has a large impact. Examples of such systems are not difficult to generate. Smoking and

asbestos exposure are both weak causes of lung cancer, but in combination they are a strong cause. A potential research question is whether there is a preferred standard model to describe systems with complimentary causes. Such a preferred model may then serve as a guide for finding a good heuristic model of strength estimation, similar to the case of causal power and the weighted ΔP rule.

5.4.4 Beyond two dichotomous causes

Another advantage of CARP is that it allows for a cleaner generalization to three or more candidate causes. For example, suppose a background cause B with capacity α_0 and a set of $\{C_i \in I\}$ binary generative candidate causes with associated capacities α_i. Then the expected probability of their conjunction is given by:

$$E[w_{T1\ldots n}|b^+, c_1^+, \ldots c_n^+] = F(\alpha_0 + \alpha_1 + \cdots + \alpha_n)$$

The expected conjunctive probability $E[w_{T1\ldots n}]$ may then be compared to the observed probability $w_{T1\ldots n}$ in order to determine the interaction between the *n* causes.

The framework also easily extends to causes measured on a continuous scale. Suppose the background is paired with the continuous cause C_1. Then predicted probability is just:

$$E\left[w_{T1(x)}|b^+, c_1 = x\right] = F(\alpha_0 + \alpha_1 x)$$

And so continuous causes may now be incorporated into the ΔP and causal power models, among others.

Danks (2014) offers an alternative algebraic approach wherein continuous causes are mapped to strengths on the [0,1] interval. Specifically, a causal strength w_i for a continuous cause C_i is defined as the expected change in the effect probability when the cause C_i increases by one unit and all other causes are at their baseline value. Clearly, then, causal strength depends on finding a natural baseline. The advantage of CARP is that α_i provides a context independent measure of strength. In addition, Danks' (2014) model largely commits to the Noisy-OR/AND representation. So another advantage of CARP is that it more easily generalizes to alternative causal models.

5.5 Causal models and the environment

5.5.1 Ecological rationality

Recall from Chapter 3 that ecological rationality asserts that behavior can only be assessed relative to the environment of its application. However, this dictum has proved hollow for the study of elemental causal induction because there have been no measurements of actual environments in which causal judgments take place. A key benefit of CARP is that it provides a representational framework by which to assess claims about relevant causal environments.

Cheng (1997), for instance, argues that the ΔP rule is not normative because it only measures covariation, not causation. CARP provides a different perspective by focusing more on plausibility instead of normativity. Trivially, the ΔP rule will be the correct model for the causal structures implied by the rule. However, we now have additional tools to unpack that claim. Above it was shown that the ΔP model is appropriate for causal contexts that are described by a uniform response curve. This implies that an identity mapping from capacity to probability: each increment of capacity produces the same increment of probability, regardless of context.

One way a ΔP system can occur is when causes are responsible for disjoint sets of outcomes. It is difficult to imagine many natural causal systems that possess such a structure. An approximate example might be found with the 1918-19 influenza pandemic. In typical populations, mortality is highest among the very young and the very old. A distinctive feature of the 1918-19 pandemic is that it primarily killed young adults, a group that generally has a low mortality rate across all populations. Thus, the ΔP model might serve as a decent model to estimate the strength of the 1918-19 flu as a cause of mortality. In contrast, the ΔP model would be less appropriate for the normal flu since it primarily kills the very young and the very old.

The weighted ΔP model is considerably more implausible as a description of a causal system. Chapter 1 showed that weighted ΔP could be represented as a parameterization on a common-effect causal graph. However, such a parameterization did not seem convincing. This is why weighted ΔP has been interpreted as an estimator rather than as a description of population relationships. The latent variable framework provides one argument for why weighted ΔP fails as a rational model. Specifically, it can be shown that there is no continuous response curve that will give a weighted ΔP latent variable model (Appendix H).

Now consider the causal power model, which assumes independent influence of the background and candidate causes. Cheng (2000) posits that people may be compelled to adopt the independence assumption because they lack sufficient information to determine how the candidate cause interacts with other causes. This assumption was incorporated into the simulation study of Section 3.12, which explored the consequences of an uncertain environment. In the simulation, the "expected" model was causal power while a disturbance term was added to create deviations from causal power. Whether independence is correct on average is, ultimately, an empirical claim. The next section speculates on how this claim might be tested.

5.5.2 Bridge to statistical models

The CARP framework can connect psychological models to the statistical literature. By doing so, it provides access to a rich set of tools that may be used to characterize actual causal environments. To see how this might work, I adapt Long's (1997) discussion of latent variable models for dichotomous outcomes. Consider the model:

$$y_k^* = \alpha_0 + \alpha_1 C_{1(k)} + \cdots + \alpha_n C_{n(k)} + \epsilon_k$$

where y_k^* is latent causal capacity and $C_{1(k)}, \ldots, C_{n(k)}$ are the observed causes associated with context *k*. The intercept α_0 represents the influence from unobserved background factors. Observed causes may be either dichotomous or continuous. The latent variable is mapped to the observed binary effect y_k using a threshold rule:

$$y_k = \begin{cases} 1 & \text{if } y_k^* > \tau \\ 0 & \text{if } y_k^* \leq \tau \end{cases}$$

where τ is the threshold. In order to identify the model, a number of parameters are set to arbitrary values. The threshold is typically set to zero as is the expectation of the error. Hence, $\tau = 0$ and $E\left[\epsilon_k \middle| C_{1(k)} \ldots C_{n(k)}\right] = 0$. To make the slope coefficients identifiable, the conditional variance of ϵ_k is assumed to equal some constant value, or $V\left[\epsilon_k \middle| C_{1(k)}, \ldots, C_{n(k)}\right] = \sigma^2$ with a finite variance $\sigma^2 < \infty$.

By assuming a specific distribution for the error it is possible to find $\Pr\left(y_k = 1 \middle| C_{1(k)}, \ldots, C_{n(k)}\right)$. Long (1997) shows that this probability can be obtained from the cumulative density function of the assumed error distribution:

$$\Pr\left(y_k = 1 \middle| C_{1(k)} \ldots C_{n(k)}\right) = F\left(\alpha_0 + \alpha_1 C_{1(k)} + \cdots + \alpha_n C_{n(k)}\right) \quad (5.5a)$$

$$w_{T1\ldots n} = F(\alpha_{T1\ldots n}) \quad (5.5b)$$

Within the statistical literature, the function $F(.)$ is referred to as the link function. For binary outcomes, the most common choices for the link function are the cumulative logistic or the cumulative normal distribution. Equation (5.5a) is the same as (5.5a), but just expressed in the CARP notation from before. It suggests that CARP response functions may be used as the link between capacity and probability. Hence, standard statistical methods can be applied to explore psychological models of causal inference.

A key question concerns the match between psychological representation and the objective structure of the environment. The first major challenge is to identify the real world causal learning contexts of interest. At present, I sidestep this first issue. Instead, I sketch potential strategies for characterizing environments once relevant data have been identified.

Hence, causal power can be expressed as a latent variable model with an exponential link function. So one strategy is to collect data from relevant causal learning environments and then fit statistical models that use an exponential link function. The statistical model could then be used to test for interactions between causes. One possibility is that interactions will be small or moderate, with positive and negative interactions tending to balance over different environments. For this set of environments, then, inference strategies that assume causal power may be well suited to the task of causal induction. The simulation study from Section 3.12 essentially conformed to this structure, and it was found that Bayesian power and weighted ΔP both performed well in estimating causal strength.

Models based on standard causal power will give poor strength estimates when causal environments are characterized by large interactions that are not symmetrically distributed around zero. Better estimates can be obtained by allowing models with causal interactions, such as the interactive causal power model from above. A disadvantage is that these models require a number

of complex steps. This seemingly makes them less plausible as descriptions of psychological processes.

Another option for dealing with interactions is to pursue a strategy similar to the one used in Section 5.4.3. On this approach, the link function is modified in order to reduce or eliminate interactions. Inference may then proceed using additive capacities only. Yet, as mentioned above, we do not want a specially tailored link function for each causal environment. Such a strategy would be more complex than finding interactive causal power. Instead, we can look for a few distinct "types" of response functions to be used for different families of causal environments. Then the judge would need to only select from a few possible strategies that match each of these types.

A rough typology of response curves can be formed based on their general shapes. For example, all non-increasing response curves could constitute one class. This class would include both the causal power model and the ΔP rule. To study inference for this, or any other class, we must know something about the response curve distribution. Section 3.12 specified the distribution a priori: causal power was chosen as the center of the class of non-increasing response curves while the variance was set arbitrarily. This a priori approach did prove useful as it allowed for general conclusions about relative estimator performance. However, it is also important to know how various estimators perform over actual environments. The next section sketches an approach for estimating response curves from empirical measurements.

5.5.3 Empirical response curves

Response curve estimation is difficult because there are a large number of unknowns, even for simple causal systems. With an unknown response curve one must estimate both capacity coefficients (the α_j's) and the response curve parameters. This section provides a rough outline on how this might proceed.

At present, I adopt the simplifying assumption that the response function is some unknown beta density. The beta density is an attractive choice since it can produce many qualitatively different curves, including ones that correspond to the ΔP rule and an approximate causal power model. Thus, the discussion below focuses on beta density estimation. In the future, however, more

general nonparametric methods should be brought to bear. For instance, kernel density estimation methods (e.g. Silverman, 1986) may be applicable to this problem.

Good targets for study will be causal systems that are both simple and high in information. These criteria can be met by a system with two observable causes, where one cause is measured on a continuous scale and the other cause is dichotomous. The continuous cause essentially serves to scale the base rate of the effect, thus allowing for evaluation of the dichotomous cause in different contexts. Suppose C_1 is the continuous cause, C_2 is a dichotomous cause, and that both are coded to be generative. Then the candidate system can be expressed as:

$$\Pr(\boldsymbol{y}|\boldsymbol{c_1}, \boldsymbol{c_2}, \alpha_0, \alpha_1, \alpha_2, \theta_1, \theta_2) = F(\alpha_0\mathbf{1} + \alpha_1\boldsymbol{c_1} + \alpha_2\boldsymbol{c_2}|\theta_1, \theta_2) \tag{5.6}$$

where $\boldsymbol{y}$ is a vector of binary outcomes, $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$ are vectors of measurements for the two causes, the α's are capacities for the background and the two candidate causes, and $F(.)$ is a beta density with shape parameters θ_1 and θ_2.

The primary goal is to estimate the response function, but it is also necessary to estimate the capacities. This is a difficult problem that requires a good deal of data to achieve reasonable precision. Figure 5.10 shows a number of attempts to estimate response curves from simulated data. For all simulations, capacities were set to $\alpha_0 = 0$, $\alpha_1 = 0.2$ and $\alpha_2 = 0.5$. The shape parameters were then varied to give different response functions. For instance, $\theta_1 = \theta_2 = 0.7$ for the response function in the first panel and $\theta_1 = \theta_2 = 1$ for the response function in the second panel. A total of 1000 observations were used to estimate each response function.

Figure 5.10 shows that a number of the estimated response curves appear close to the true curves, while several others miss badly. Figure 5.11 plots the estimated capacities and overlays the population values. The coefficient estimates are generally grouped around their true values, though there is considerable error for the α_2 coefficient.

Despite the relatively large number of simulated observations, the estimation algorithm still produces sizeable errors. It remains an open question whether appropriate data and statistical methods exist so that causal systems can be estimated with sufficient precision. The outlook is somewhat daunting due to the large number of unknowns for even simple causal systems. However, even very coarse estimates could prove useful to the study of causal learning. For instance, it may only be possible to determine that the objective response curve comes from some broad class, such as the class of all increasing response curves. This is still valuable information

as it can be used to study inference in the context of model uncertainty. Objective measurements from the environment would inform model uncertainty instead of it being determined completely a priori, as was done in Section 3.12.

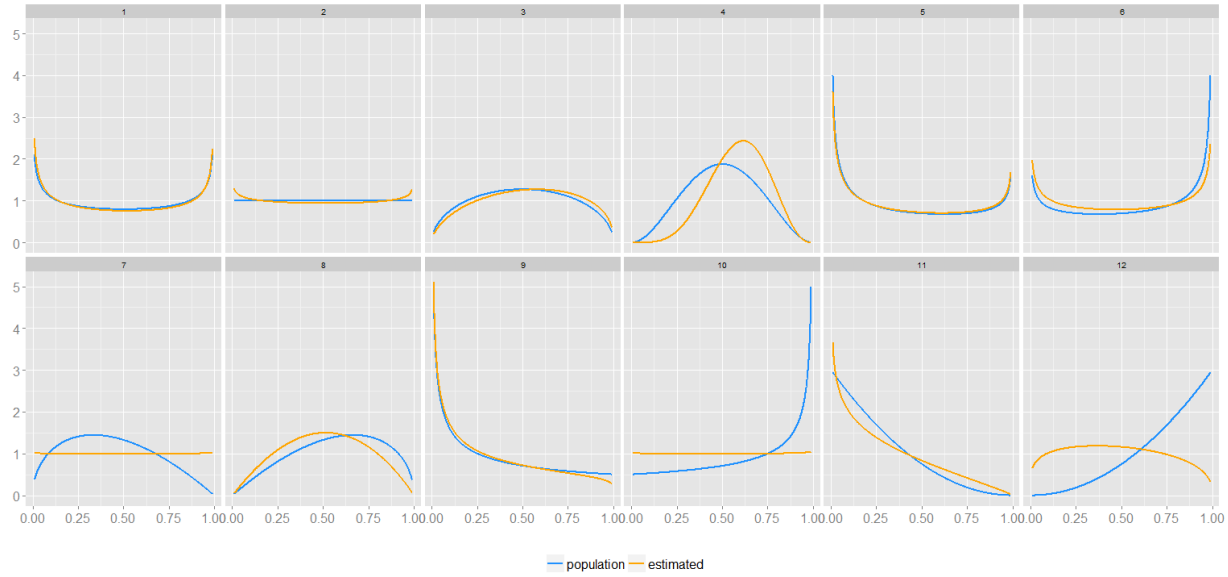


Figure 5.10. Twelve response curves (blue) and estimated response curves (orange) from a simulation study. The population curves were all chosen from the beta family. The population relationship involved three unknown causal capacities and two unknown shape parameters, as shown in equation (5.6). Details of the estimation procedure are given in Appendix I.

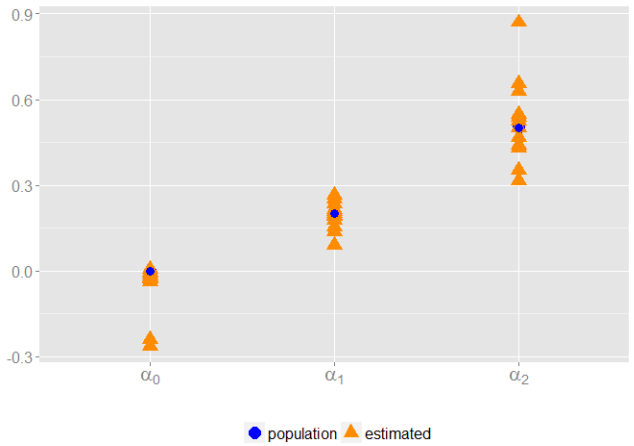


Figure 5.11. Causal capacity estimates for the twelve response functions shown in Figure 5.10. True parameter values were $\alpha_0 = 0$, $\alpha_1 = 0.2$, and $\alpha_2 = 0.5$ (shown in blue). Details of the estimation procedure are in Appendix I.

Appendix I outlines a couple of potential methods that can be applied to the causal systems estimation problem. It includes the algorithm used for Figure 5.10 and Figure 5.11, as well as a Bayesian method that uses the Metropolis algorithm.

5.6 Human learning of response functions

Earlier, it was proposed that response curves could be categorized into a few general classes and a heuristic rule could be applied for each class. Such a strategy can return decent judgments with minimal computation. Section 3.12, for example, showed that weighted ΔP is a good heuristic when models are drawn from the class of non-increasing response curves.

A heuristic judgment strategy seems adaptive when learning information is limited. But what about when there is a wealth of information? Kahneman (2011) asserts that intuitive expertise can be developed when 1) the environment is sufficiently regular and 2) there is an opportunity for repeated practice with fast, high quality feedback. Indeed, people have demonstrated a remarkable ability to learn very fine associations when these two criteria are met. It should seem possible, then, for people to develop expertise for particular causal environments when given sufficient learning opportunities. This section uses the CARP framework to outline a possible mechanism for how such learning could proceed.

To begin, suppose a system with a binary effect E and two generative causes with C_1 continuous and C_2 binary. The binary variables are coded 0 and 1 for their absence or occurrence, respectively. For simplicity, assume that the background capacity is zero, or $\alpha_0 = 0$.

Imagine that repeated experience has allowed the judge to learn probabilities that are close to their population values for certain combinations of causes. Specifically, suppose that $c_1 = x$ and $c_1 = z$ are two levels of the continuous cause where $x < z$, and that the judge has learned the probabilities:

$$\begin{aligned} \Pr(e = 1 | c_1 = x, c_2 = 1) &= F(\alpha_1 x + \alpha_2) \\ \Pr(e = 1 | c_1 = z, c_2 = 1) &= F(\alpha_1 z + \alpha_2) \end{aligned} \tag{5.7}$$

where $F(.)$ is some unknown response function. Note that $F(\alpha_1 x + \alpha_2) \leq F(\alpha_1 z + \alpha_2)$ since response functions are increasing. Now suppose that the judge is interested in predicting the

probability for the novel causal combination of $c_1 = y$ and $c_2 = 1$ where $x < y < z$. So the judge would like to know $F(\alpha_1 y + \alpha_2)$.

Without knowing the response function and the causal capacities, is it possible to make a principled prediction? If the response function is smooth, and not too unusual in shape, then a linear approximation may do the job. The secant approximation is given by:

$$F(\alpha_1 y + \alpha_2) \approx F(\alpha_1 x + \alpha_2) + \frac{F(\alpha_1 z + \alpha_2) - F(\alpha_1 x + \alpha_2)}{\alpha_1 z + \alpha_2 - (\alpha_1 x + \alpha_2)} \times [(\alpha_1 y + \alpha_2) - (\alpha_1 x + \alpha_2)]$$

Cancelling produces:

$$F(\alpha_1 y + \alpha_2) \approx F(\alpha_1 x + \alpha_2) + \left(\frac{y - x}{z - x}\right) \times [F(\alpha_1 z + \alpha_2) - F(\alpha_1 x + \alpha_2)] \qquad (5.8)$$

Note that all quantities in (5.8) are observable! The *x*, *y* and *z* are just the different observed levels of cause C_1 while $F(\alpha_1 x + \alpha_2)$ and $F(\alpha_1 z + \alpha_2)$ are the observed probabilities. Working through a bit more algebra:

$$\begin{aligned} F(\alpha_1 y + \alpha_2) &\approx \left(\frac{y - x}{z - x}\right) \times F(\alpha_1 z + \alpha_2) + \left[1 - \left(\frac{y - x}{z - x}\right)\right] \times F(\alpha_1 x + \alpha_2) \\ &= k \times F(\alpha_1 z + \alpha_2) + (1 - k) \times F(\alpha_1 x + \alpha_2) \end{aligned} \qquad (5.9)$$

with $k = \left(\frac{y-x}{z-x}\right)$. Following the theme from earlier chapters, the prediction is found from a linear combination of observed probabilities.

The interpolation model of (5.9) nicely lends itself to psychological interpretation. Exemplar-based theories of categorization assert that novel instances are classified based on their similarity to previously known instances in memory (Nosofsky, 1986, 1988). The interpolation model suggests a similar mechanism, though now the prediction is quantitative. Specifically, when the judge encounters a novel causal combination, they may retrieve exemplars from memory in which the outcome probability is known. The exemplars can then be weighted according to their similarity to the novel case. This is just what equation (5.9) does when similarity is defined as the Euclidean distance between the novel case and the known exemplars. When the novel case is close to $c_1 = x$, then $k \to 0$ and the $F(\alpha_1 x + \alpha_2)$ probability gets most of the weight. Likewise, when

the novel case is close to $c_1 = z$, then $k \to 1$ and the $F(\alpha_1 z + \alpha_2)$ probability gets most of the weight.

Secant approximation is a simple method that uses known probabilities to make predictions. These predictions will clearly improve as the judge learns more of the true probabilities. This learning can be interpreted as building up the response function through experience. Figure 5.12 shows how the process develops.

In panel A of Figure 5.12, the judge has learned population probabilities for four causal combinations. The secant lines connect the adjacent probabilities to give a piecewise linear approximation of the response function. The response curve is just the derivative of the response function, so the piecewise linear approximation implies a step function approximation of the response curve, which is shown in panel B. Naturally, as the judge learns more population probabilities the approximation improves. Panel C shows the response function approximation when seven probabilities are known and panel D displays the corresponding step function.

With secant approximation, predictions can be made using only observed probabilities. So is there any reason to form a representation of the response function? Consider again the case in which the two probabilities from (5.7) are known, but now suppose the judge is interested in the novel combination of $c_1 = y$ and $c_2 = 0$. That is, the judge wants to predict $\Pr(e = 1 | c_1 = y, c_2 = 0) = F(\alpha_1 y)$. In the first example, all unknown α's cancelled in the approximation formula. But for the current example, the unknown $\alpha's$ will no longer cancel, so the secant formula can no longer be used. The problem is that the novel combination differs from the known combinations on both the C_1 and the C_2 causal dimensions. It is for these types of situations that the response function is desirable.

When the response function is known, capacities can be inferred from probabilities. Predictions can then be made for any novel combination of causes. This is precisely the utility of capacities argued for by Cartwright (1989), as was discussed in the introduction to this chapter. In summary, investing resources to represent the response function may be sensible since it allows the judge to infer causal capacities, which may then be used for prediction in novel contexts.

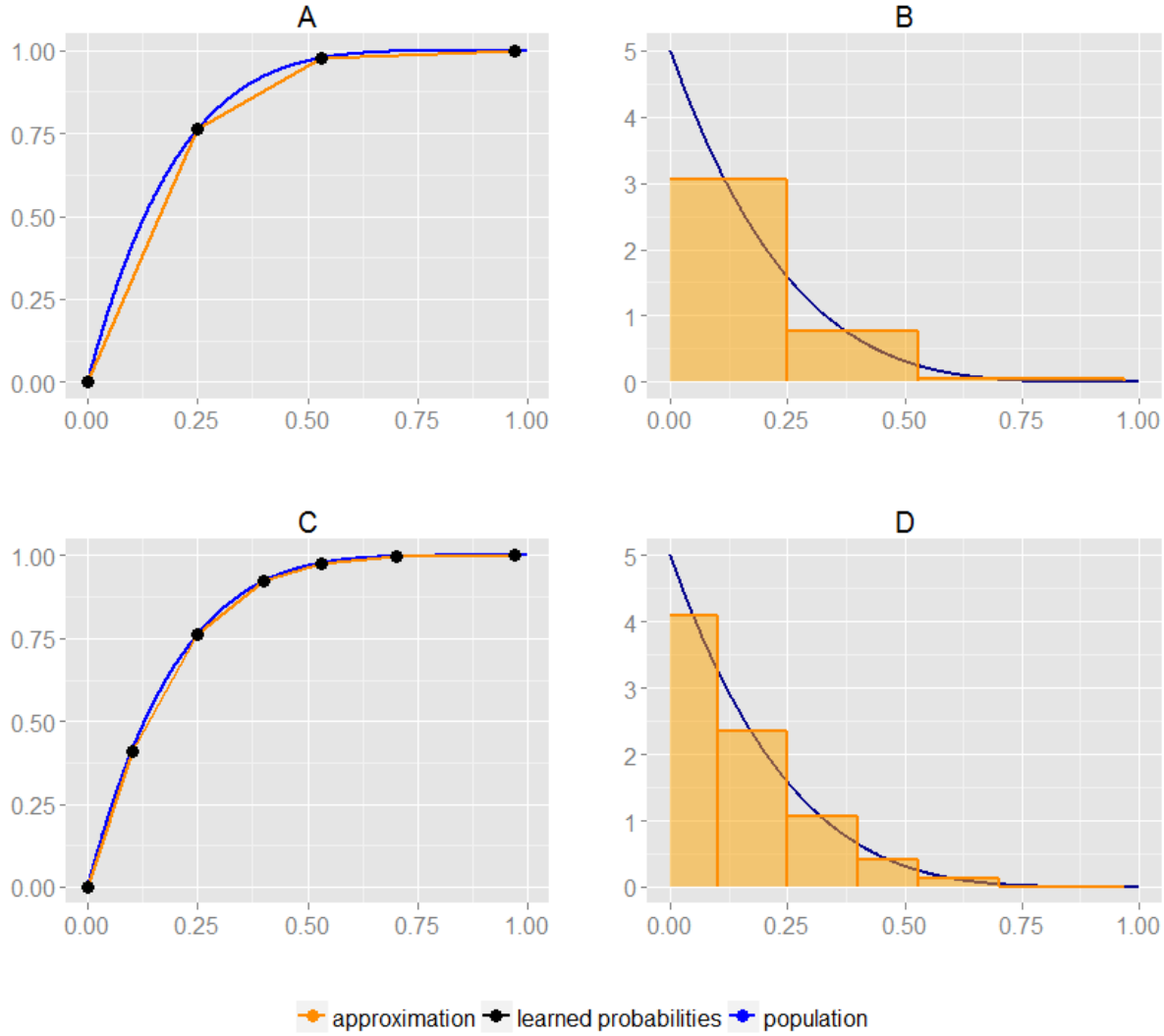


Figure 5.12. Piecewise linear approximations of a cumulative response function (left panels) and the implied step function approximations of the response curve (right panels).

A: The four black points represent true probabilities that have been learned to high precision, so they all lie on the true response function shown in blue. The secant lines give a piecewise linear approximation, shown in orange.

B: True response curve (probability density) is shown in blue. The piecewise linear function from **A** implies a step function approximation of the response curve.

C: Seven known probabilities gives a better approximation to the cumulative response function.

D: The step function approximation from the seven probabilities in **C**.

5.7 Conceptual applications

Previous sections have shown that CARP can provide a lens through which to view the external world of causal environments. Now the lens is turned inward. If one accepts CARP as an approximate description of causal intuitions, then the formalism can be used to study these

intuitions. This section assumes CARP as a descriptive theory. Namely, it assumes that people's psychological representations can be described using the proposed latent variable framework. Whether this assumption is justified is a topic of future study. However, its tentative acceptance allows one to address various conceptual issues in the study of causal learning.

5.7.1 Causal probe question

Chapter 1 discussed the disagreement about the wording of the question used for causal strength judgments. Recall a criticism of the standard causal probe is that it is ambiguous with regards to the context in which the strength judgment is made. To remedy this deficiency, a number of researchers use questions that are explicit about the context. For instance, Buehner et al. (2003) favor a counterfactual wording in which:

> The context (before the intervention) we chose for estimating the extent of generative influence was one in which *e* never occurred, and the context for estimating the extent of preventive influence was one in which *e* always occurred. In these contexts, because alternative influences of *e* of like kind as the candidate (generative or preventive) are counterfactually removed, the influence of the candidate should manifest itself without contamination. That is, the estimated frequency of *e* in the (counterfactual) presence of the generative cause should reflect the strength of this cause alone; the same holds for the estimated frequency of $\neg e$ as a measure of the preventive strength of the candidate. *There are no simpler or clearer contexts under which to manifest the strength of a candidate causal relation* [emphasis added]. (Buehner et al., 2003, p. 1128)

Subsequent researchers using counterfactual or suppositional wording have followed the above advice. One example is found in the generative causal query of Lu et al. (2008, p. 962):

> Suppose that there is a sample of 100 DNA strands and that the gene is OFF in all those DNA strands. If these 100 strands were exposed to the protein, in how many of them would the gene be TURNED ON?

Moreover, Lu et al. (2008) argue that the critical feature of this wording is not its precision, but that it measures strength in a context where no other cause is producing the effect.

It is important to be clear regarding what is intended by the counterfactual or the suppositional wording. For generative scenarios, removing all alternative influences is meant to preclude the possibility of some other cause producing the effect. Thus, for the generative causal probe, the probability of the effect will equal zero in the reference context. In other words, participants should imagine a situation in which it is impossible for the effect to occur before introduction of the candidate cause.

Though the intended meaning of the causal probe is made clear in the discussions of Buehner et al. (2003) and Lu et al. (2008), the wording of the question itself does not strictly convey this precise meaning. A seemingly legitimate interpretation of the above question is that the effect could have occurred, but that it just failed to for the particular hypothetical sample under consideration. Such an outcome is consistent with a small positive probability of the effect for each of the hypothetical trials. In fact, this alternative interpretation would often seem more reasonable for a number of domains.

A bit of reflection suggests intuitive reasons why the intended hypothetical context is problematic, and CARP can make these intuitions precise. The focus will be on the generative case, though with slight modification most of the criticisms will hold for preventive scenarios as well. The key problem is that contexts in which the effect never occurs are often highly ambiguous. Indeed, for probabilistic causation they appear more ambiguous than contexts in which the effect occurs with some positive probability.

Consider the counterfactual or suppositional context where no other cause is producing the effect. To be sure, certain examples will precisely satisfy this hypothetical constraint. An archetypal one is found in Newton's first law of motion: an object at rest will stay at rest unless some external force acts upon it. Hence, the probability of the effect (motion) will be zero when all causes (forces) are removed. Key features of this example are that it refers to a simple physical system with (near) deterministic causality.

The situation becomes much murkier when one considers causal scenarios that are complex and probabilistic, which is typical of those used in causal learning experiments. For these scenarios, removing alternative causes is not simply an all-or-none proposition. Instead, one must scale down alternative causes to a level at which they are no longer sufficient to generate the effect.

To make these issues concrete, consider the effect of "high school graduation". There are many causal factors that influence whether a high school student graduates, with a large proportion

mediated through the general constructs of "motivation" and "ability". For example, improving student nutrition will improve both as it increases energy levels and also boosts several cognitive faculties (Pollitt, Cueto, & Jacoby, 1998), so we should also expect it to increase graduation rates.

Now suppose we observe that some educational intervention, such as a free school meal program, improves the graduation rate at a given high school from 60 to 70 percent. Or equivalently, the probability that a randomly sampled student graduates from this high school improves from 0.6 to 0.7. To judge causal strength using the advice from Buehner et al. (2003), one should imagine a hypothetical student for which graduation does not (or will not) occur, and then imagine how the probability changes once the cause is introduced. So for the school example, we must remove the motivational and ability factors that allow the hypothetical student to graduate, and then imagine the influence of the intervention. The problem is that many different hypothetical students come to mind who satisfy this requirement. We can imagine students who just miss graduation by failing to pass one or two classes, as well students who drop out in their first year of high school.

Some may object to the high school example since it evokes an abundance of background knowledge. Yet the same type of criticism can be made for virtually any probabilistic causal scenario. Consider Lu et al.'s (2008) gene activation scenario from above. Gene expression is a complex multi-stage process that can be influenced to varying degree at any of these stages (Russell, 2003). A gene may fail to activate in an otherwise healthy cell because there is currently no use for the protein that the gene would produce. Alternatively, the gene may fail to be active because the cell is highly stressed with few metabolic resources. Though both backgrounds result in no activity, these are two highly different reference states that should respond quite differently to the introduction of some candidate cause of interest.

The CARP framework can be used to show why a hypothetical context with a zero effect probability is especially ambiguous. To see how, assume an exponential response function $F(.)$ and a generative candidate cause C_1 with associated capacity α_1. If we follow Buehner et al. (2003), the causal strength of C_1 will be evaluated in a context with the probability of the effect equal to zero, or for capacities α_i such that $F(\alpha_i) = 0$. With an exponential response function this condition is met for all $\alpha_i \leq 0$. Now if the judge is able to scale down the alternative causes or scale up preventive factors so that $\alpha_i = 0$ exactly, then introducing C_1 to this hypothetical context results in $F(0 + \alpha_1) = F(\alpha_1)$. This gives the causal power prediction. However, the judge might

imagine a more conservative reference context. This can be represented in CARP with a negative capacity $\alpha_h < 0$, which still satisfies the zero probability condition $F(\alpha_h) = 0$. The judged causal strength then becomes $F(\alpha_h + \alpha_1)$, a quantity that does not equal causal power.

The left panel of Figure 5.13 depicts the problem of an ambiguous reference context. When a reference capacity of $\alpha_i = 0$ is used, causal power is given by the pink plus the dark red area. But when a more conservative $\alpha_h = -3/4$ is used for the reference context, the strength estimate is given instead by the pink area only.

A second criticism of Buehner et al.'s (2003) recommend probe is that it makes sense only for a subset of causal models. Namely, the wording assumes that a context with zero effect probability exists and is meaningful. Yet it is plainly evident that many causal domains do not fulfill this assumption.

Two general types of causal systems will not conform to the zero effect assumption. First, for many causal domains a zero effect probability is not within the natural range of outcomes. For an example, consider the probability that a neuron fires an action potential within some fixed time

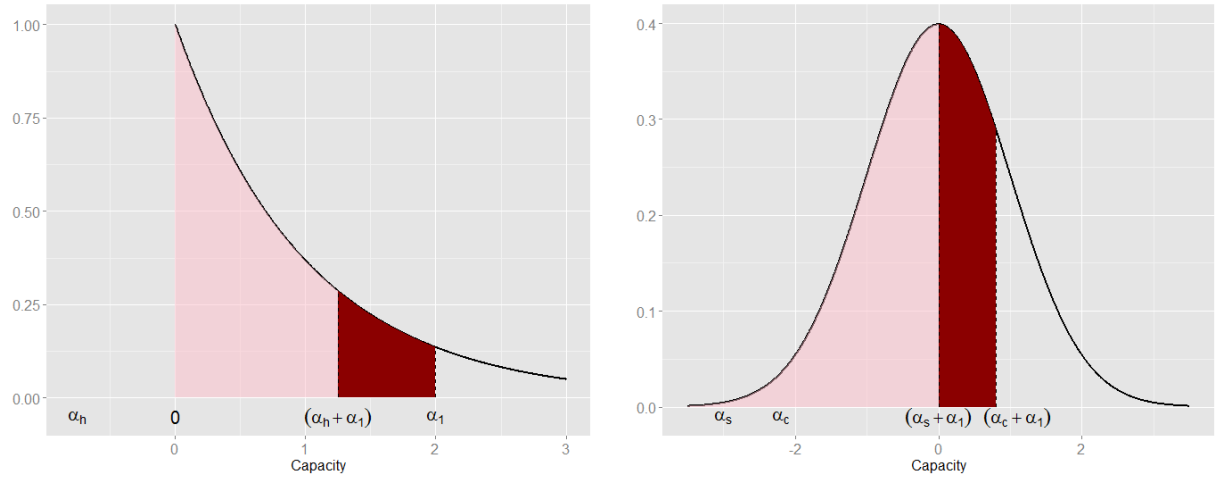


Figure 5.13. Response curves that represent heterogeneity of reference points for zero or near-zero probabilities. In both panels, a cause C_1 with capacity α_1 is evaluated using different reference contexts.

A: The exponential response curve gives causal power when the reference context capacity is zero. Causal power is shown by the pink area plus the red area. When reference capacity is below zero, the judgment does not equal causal power. The judgment for a reference capacity of $\alpha_h = -0.75$ is shown by the pink area only.

B: For finite capacities, a standard normal response curve does not allow for a reference probability of zero. Instead, the judge may choose capacities that give a "small" reference probability. Two choices are $\alpha_s = -3$ and $\alpha_c = -2.2$, which give two different causal strength judgments shown by the pink and the pink plus red areas, respectively.

interval. This probability can be modulated up or down with excitatory or inhibitory neurotransmitters. However, a zero probability of firing is not within the natural range, and in fact, would probably only occur if the neuron was dead.

A second type of problem occurs if the probability can be made close, but not exactly equal to zero. For instance, try to imagine a group of adults who will definitely not die of cardiac arrest over some duration. Such a thought experiment will fail since there is always some probability of cardiac arrest, even in otherwise healthy individuals (Sen-Chowdhry & McKenna, 2006). Nonetheless, through creative imagination we may contrive ever-more elaborate contexts to safeguard these hypothetical individuals. We might imagine that they are young, healthy, have no family history of heart problems, were recently examined for congenital defects, and so on. So while a true zero probability may be unattainable, it may be possible to get closer and closer with more elaborate hypotheticals.

Once again, there is a risk that people will vary the extent of the reference context hypothetical, though now the problem may be more severe. With a true zero probability there is at least the potential for respondents to all form the same reference context. Without a zero probability, respondents will be left to decide what is close to zero, and this freedom should contribute additional variation in the responses.

Once again, the CARP framework allows for a clear demonstration of the problem. The right panel of Figure 5.13 above shows a standard normal density for the response curve. Accordingly, the response function has positive support over the entire real number line, and a zero probability is not possible for any finite capacity. Suppose a candidate cause C_1 and with an associated capacity $\alpha_1 = 3$. When asked to imagine a zero probability reference context, participants may adopt different baselines. This is shown in the figure with capacities $\alpha_s = -3$ and $\alpha_c = -2.2$, which respectively give reference probabilities of $F(\alpha_s) \approx .001$ and $F(\alpha_c) \approx .01$. So both reference contexts are close to zero, though the α_s context is an order of magnitude smaller in probability. The natural consequence of these varied reference contexts is differing causal strength estimates. For the more conservative context, the estimate is $F(\alpha_s + \alpha_1) = 0.5$, which is represented by the pink area in the left panel. And for the more permissive context, the estimate is $F(\alpha_c + \alpha_1) \approx 0.79$, which is shown by the pink area plus the red area.

Thus, the reference probability will necessarily be positive for some causal systems while Buehner et al.'s (2003) recommend probe tries to deny this reality. Participants are then left to

determine a probability that is sufficiently close to zero. Compounding the problem, this reference probability would seem to differ based on whether a counterfactual or suppositional context is used. The suppositional context is one in which you can predict with certainty that the effect will not occur. This implies more extreme antecedent factors than those associated with a counterfactual outcome, for which one can examine a particular case and imagine why the outcome failed to occur. Returning to the high school example, to know for certain that a student will not graduate suggests that they come from a highly disadvantaged background. In contrast, the counterfactual student who does not graduate is more permissive as it extends to students who just miss graduation by a couple of credits.

In summary, Buehner et al. (2003) are right to emphasize the paramount importance of invoking a consistent context for causal strength evaluations. Doing so has key implications for both *intra* and *inter* individual judgments. Within a respondent, adopting a constant reference context should produce consistent ordinal strength judgments. It is also important to encourage different respondents to adopt a similar reference context. As has been discussed in previous chapters, it is standard to use group statistics when reporting strength judgments. If different people assume different reference contexts, then it will increase the variance of these measures and make model comparison more difficult.

While Buehner et al. (2003) highlight a key concern, their proposed amendments may make the reference context even more ambiguous. In fact, recent research supports this hypothesis: Shou and Smithson (2015) found that "predictive" causal questions, like the one above from Lu et al. (2008), produced more variability in responses relative to the more standard causal probes.

CARP proves a valuable tool in demonstrating potential problems in the causal strength probe. Might it also help point towards a better question construction? Under the CARP representation of causal power, with an exponential response function, only the zero probability has a many-to-one mapping of capacities to probability as all $\alpha_i \leq 0$ return $F(\alpha_i) = 0$. Yet for any positive capacity $\alpha_i > 0$, the function $F(\alpha_i)$ is one-to-one. Intuitively, a positive probability reference context may evoke a more specific set of instances. This will be especially true when the probability is within the natural range for the domain of interest. Returning to the graduation example, it is much easier to imagine a high school with a sixty percent graduation rate than it is to imagine one with graduation rate of zero.

Thus, it should be preferable to use a causal probe that invokes a positive probability reference context. A good general strategy could be to use a small positive reference probability for causal scenarios in which a zero probability sensible. And for scenarios in which a zero probability is outside the natural range of outcomes, the causal question could invoke a reference probability at the lower end of the natural range.

Cartwright observes in her book that, "…fundamental laws are laws about distinct 'atomic' causes and their separate effects; but when causes occur in nature they occur not separately, but in combination," (Cartwright, 1989, p. 175). This passage highlights a central tension in the attempt to measure causal intuitions. On the one hand, the cleanest formal representation is one that isolates the cause. But on the other, isolated causes do not typically present themselves to everyday experience. The upshot is that intuitions are unlikely to correspond to isolated causes, and so measurements that attempt to solicit such intuitions will most likely be misconstrued.

5.7.2 Axiomatization of CARP

An axiomatic account begins with a few basic assumptions (axioms) and then derives a number of results from these assumptions. Axiomatic measurement theory seeks to derive properties of numerical assignment from a set of foundational assumptions (Krantz, Luce, Suppes, & Tversky, 1971). The approach has been used to explicitly establish and systematize properties of numerical scales for various areas of inquiry. For a particular scientific domain, the basic assumptions will concern how the units of analysis relate to and combine with one another.

In motivating the latent variable model, a number of basic assumptions were made about combinations of causes and their relation to probability of the effect. Namely, CARP assumes that causes combined additively in their capacities and that capacity was mapped to probability by a strictly increasing response function $F(.)$. Additional assumptions about causal strength then allowed for derivation of the ΔP and causal power models. Based on these assumptions, it would seem that there is potential to give an axiomatic account of these models. In fact, an axiomatization is possible, as is shown in Miyamoto (forthcoming).

5.8 Summary

In earlier chapters, only two candidates for the data generating process were considered: the ΔP rule and causal power. And for the majority of this text, causal power has been treated as the "true" underlying model. Even with this simplifying assumption, it has been quite difficult to choose the best model to explain human causal judgment, be it a Bayesian model of causal power or the weighted ΔP rule. Results from this chapter would seem to only complicate the question, as now there is an infinite set of potential data generating models.

The causal-power-as-truth story is simple, though it comes with its price. Many difficulties from previous chapters were largely due to the absence of an account of the objective causal environment. For instance, Chapters 2 and 4 showed how some description of the relevant environment is crucial in making the case for Bayesian models as optimal. The CARP framework provides a bridge between models of human causal inference and standard statistical methods. As such, it provides conceptual machinery that can be used to measure the actual environments in which judgments take place. Considerable work is necessary to identify the relevant environments and the statistical tools that may be used to measure them. Hopefully, such work will be a fruitful area of future research.

Chapter 6.

Conclusions

> Maybe you can't afford the ideal [solution], but if we can approximate it in a certain way, you can get 98 percent of the benefit with 1 percent of the computation.
>
> –Jeff Dean, the "Chuck Norris" of the Internet (from Slate's "The Optimizer").

6.1 Contributions of the rational Bayesian approach

Our survey of the field of causal learning has covered considerable territory. It has reviewed several competing research traditions and many models that have arisen from them. A central focus has been on the rational Bayesian approach. Much of the discussion has been critical, but it is important to emphasize the many laudable aspects of this program. Early rational models of causal learning essentially assumed that true probabilities could be gleaned from learning data. The Bayesian perspective provided a key corrective: uncertainty is fundamental to inference. No model can be rational unless it accounts for uncertainty. Hence, one major contribution of the Bayesian approach has been to identify a key oversight in the original analysis.

Bayesian analysis offers a powerful set of tools to characterize problems involving inductive inference. It provides a precise mathematical framework for model representation. This enforces a uniformity on psychological theory and lays bare underlying assumptions, both desiderata of the scientific enterprise. Further, the Bayesian approach can deliver optimal solutions for certain inferential problems. Even if one rejects Bayesian cognition, the ideal solution can function as a useful reference point by which to evaluate heuristic models, as was done in Chapter 3.

As some philosophers have observed, a theory that does not provide true causal explanations can still be useful as a descriptive theory (Danks, 2008). That is, a theory can serve to organize a large assortment of phenomena even if it fails to pick out the underlying causal mechanisms. Might the Bayesian approach perform such role by providing a standard and uniform account of human cognition? At present, it is much too preliminary to decide. It may turn out that a hodge-podge of idiosyncratic assumptions will be necessary to explain the vast diversity of human behavior. This

outcome may represent limitations of the Bayesian formalism. Or, as some researchers believe, it may simply reflect the underlying cognitive reality. The mind might be a kluge (Marcus, 2008), an assortment of mental modules (Barkow, Cosmides, & Tooby, 1992), or an adaptive toolbox(Gigerenzer & Selten, 2001).

6.2 The cost of computation

Bayesian methods have made strong contributions to cognitive science and they should continue to do so. Yet its clean formalism and promise of optimality can be beguiling. As with any scientific endeavor, one must proceed with a healthy degree of skepticism. A major goal of this dissertation has been to assume the skeptic's post and shine a critical light on Bayesian models of cognition.

The primary standard that has been used to evaluate Bayesian models has been empirical. Chapters 3 and 4 showed that the one parameter weighted ΔP model gives a better description of judgments than leading Bayesian models of parameter estimation. However, there will always be a limit to what may be distinguished empirically. The end of Chapter 4 presented a two-stage Bayesian inference model that gives predictions that closely mirror weighted ΔP. Further, theory tells us that the Bayesian model has the potential to be optimal while there are no similar results for weighted ΔP. So why not choose the Bayesian model?

For a Bayesian solution to be truly optimal, the problem must be narrowly demarcated. All costs need to be explicitly defined in terms of a loss function. Certainly, such a rigorous quantitative approach is desirable for a theory of cognition. Yet even more certainly, rigor should not come at the expense of veracity. The cognitive costs for maintaining and operating over mental representations are very difficult to measure. But this does not make them any less real. Indeed, the underlying metabolic costs may be quite significant. Lennie (2003) estimates the metabolic cost of individual neuronal spikes, and from this estimate he finds that very few neurons can be active concurrently in the human cortex—possibly fewer than 1%. For this reason, among others, a number of researchers conjecture that the brain must use sparse encoding of representations (Bowers, 2009; Lennie, 2003). Accordingly, Markman and Otto (2011) contend that any definition of optimal behavior should also include some account of energy expenditure. Of course, these ideas are by no means new. Simon (1955), for instance, famously argued that resource constraints need to be incorporated into theories of rational behavior.

Bayesian advocates have come to recognize the importance of computational cost. Chapter 2 mentioned the method of "resource rational analysis" used to construct rational process models, which are meant to form a bridge between computational and algorithmic levels (Griffiths et al., 2015). Though this is a move in the right direction, the approach appears overly constrained. The main limitation is the requirement that rational process models converge to the objective function in the limit. As a consequence, a minimum overhead is built-in to the representational assumptions. To date, the objective function has always been Bayesian. Thus, rational process models assume that people form representations of a prior distribution and a likelihood function, that these are combined according to Bayes rule, and that there is some mechanism that samples from the posterior distribution. All potential rational process models include these representational commitments, so they are effectively omitted from the ledger of cognitive costs. The result is a restricted definition of computational cost, defined only in terms of sampling.

In Bayesian statistics, the posterior distribution is useful for many reasons. Applications include finding posterior confidence intervals, the posterior predictive distribution, and posterior distributions for arbitrary functions of the modeled parameters. Yet in rational process models the use of the posterior quite limited. It is used only to take a few samples of the relevant parameter in order to estimate its value. As Vul et al. (2014) note, approximations can be quite poor when the number of samples from the posterior is small, leading to strong deviations from the Bayesian solution. In sum, rational process models pay a steep price in representational commitments while returning a narrow benefit, which may be of dubious value.

Resource rational analysis would never produce the weighted ΔP model since it is fundamentally non-Bayesian. That is, uncertainty is not represented with probability distributions and belief is not updated according to Bayes rule. Though it is not Bayesian, weighted ΔP does a very good job of approximating the Bayesian solution. This was shown in the simulation from Section 3.6 in which weighted ΔP was almost as good as the normative Bayesian model, with the two models agreeing over 90 percent of the time. Any rational process model that relies on a small posterior sample will almost certainly be inferior. Further, weighted ΔP achieves this performance at a fraction of the representational and computational commitments relative to Bayesian models.

There should be little controversy regarding the claim that the weighted ΔP model entails lower computational cost relative to Bayesian models. In other comparisons, especially among competing algorithmic models, this assessment will be more difficult. This is why Danks and

Eberhardt (2011) cite the need for an account of algorithmic rationality. One starting point might use relative processing costs from computer science. For instance, we know that it costs more to represent a distribution than it does to represent a single value. Similarly, the computer operation of multiplication is more costly than addition. Computer science can tell us precisely how much more expensive in terms of the best known algorithms used to execute these functions. These findings could serve as initial assumptions, which may eventually be replaced as our understanding of psychology and computer science improves. Such an approach is essentially the same as resource rational analysis. The one major difference is that Griffiths et al. (2015) only consider the class of algorithms that are used to approximate a posterior probability distribution.

In short, it is time to bring "computation" to the forefront of the computational level and think less like mathematicians and more like computer scientists. Instead of searching for the ideal solution, we may instead search for solutions that are close to ideal while saving massively on computation time. Perhaps models that fulfill these dual objectives can then be known as "Jeff Dean rational".

6.3 Future directions

What are people doing when they form judgments of causal strength? How do they use evidence to form these judgments? Substantial progress has been made on these questions in the 50-plus years since the first studies on contingency judgment. Most of this progress has been conceptual. Cognitive models have become increasingly refined, spurred in part through the evolution of formal frameworks in statistics and computer science. Yet at the same time, most contemporary experiments remain quite similar to those used in early research. In part, this is because the central research questions have remained largely unchanged. Nonetheless, improved measurement is one pre-requisite for the advancement of science.

Of course, establishing a refined characterization of the research problem often leads to novel predictions and to improvements in experimental method. While this dissertation has focused primarily on theory development, much of this work has been a theoretical means to empirical ends. Some predictive payoffs have been immediate. Chapter 3 reinterpreted the weighted ΔP model as an estimator of causal power. By doing so, it became apparent that weighted ΔP had previously been misspecified for preventive causes. The correct specification, which changes the

focal event from effect present to effect absent, demonstrates a profound improvement in fit to experimental data.

The revised weighted ΔP specification also suggests additional measures that might better distinguish models. If effect absent trials are focal for preventive causes, one should expect that they will receive more attention during the learning task. Time-per-trial could serve as a proxy for attention, and this measure is easily obtained since nearly all experiments are now administered by computer.

Analysis of weighted ΔP also contributed to the discovery of the deterministic bias. As shown in Chapter 4, the deterministic bias was demonstrated by a majority, though not all, participants in certain key conditions. This result underscores the importance of staying alert to individual differences in judgment strategies. The deterministic bias was probably concealed in previous studies by the use of group averages.

Somewhat curiously, individual differences did receive attention in early studies of contingency judgment. For instance, Ward and Jenkins (1965, p. 234) stated that, "It became clear in preliminary experiments that different subjects adopt distinctly different bases of judgment. Hence, group averages provided little useful information." So instead of using group statistics, Ward and Jenkins (1965) attempted to classify each person as following one of four distinct judgment rules.

Going forward, the study of causal learning would benefit from a renewed focus on the characterization of heterogeneous response strategies. Performing such analyses requires repeated measurements across multiple conditions. This presents a challenge, especially for sequential presentation formats that use many trials per condition. One solution could be to measure participants on multiple occasions, which should mitigate the influence of fatigue.

Thus, we see several examples of how theory development has led to novel predictions while also suggesting improvements in method. Another example is found with the power PC model, which has influenced several aspects of experimental design. In order to ensure that the power PC model assumptions are met, causal learning stimulus materials have become increasingly patterned after hypothetical scientific studies. This is true of the experiments reported in Chapter 4, which were essentially replications of previous studies. There are several reasons why a hypothetical scientific context is a good choice for testing the power PC theory. First, it limits the influence of background knowledge on judgments. More importantly, most college students (the typical

participants) should know that scientific studies take special care to avoid confounding. Accordingly, participants should feel justified in believing that the candidate cause is statistically independent of background factors, a critical assumption of the power PC model.

To summarize, the power PC model explicitly described the type of environment for which it is appropriate, and researchers subsequently attempted to make stimulus scenarios resemble this environment. The CARP framework of Chapter 5 aspires to reverse this arrow of influence. That is, it aims to allow for the measurement of actual causal environments which can then inform experimental design and analysis. For example, we might be interested in causes of some disease. The right data would allow researchers to characterize how genetic, behavioral, and environmental causes influence the probability of occurrence. On the CARP approach, this would involve estimating a response curve under the assumption of additive causal capacities. The very same data could be used to construct stimulus scenarios with which to train participants about the causes. Comparisons could then be made between participant judgments and model predictions for various novel causal combinations. An interesting extension might be to test judgments of experts versus those of laypeople. In the above example, experts would be medical professionals with clinical experience predicting the disease.

Increased control in cognitive science is often achieved by stripping away context. For example, classic research in categorization uses sparse, abstract stimuli such as geometric objects (e.g. Medin & Schaffer, 1978; Nosofsky, 1986). This is sensible strategy for several reasons. As just mentioned, abstract materials limit the potentially contaminating role of background knowledge. And sparseness ensures that participants are attending to the manipulated variables of interest. In short, the goal is to isolate the cognitive process and minimize measurement error. The risk, of course, is that it becomes increasingly difficult to generalize the findings to real-world behavior. In the study of causal learning, experiments have, if anything, only become more abstract. To combat this threat to ecological validity, future research will need to use scenarios that evoke real world contexts. The CARP framework provides one avenue by which to introduce the reality of the external world while, at the same time, hopefully also organizing its messiness.

Appendix A.

Power PC derivations

A.1 Generative causal power

This section notes difficulties with one of the original assumptions from Cheng's (1997) power PC Theory. Instead of refining the assumption, which could prove difficult, I propose an alternative derivation that obviates its use.

Let *C* be the observed candidate cause, *B* be a composite of all unobserved causal factors that are on net generative, and *E* will represent the effect. All variables are binary. Recall the four primary assumptions of causal power are:

1) *B* and *C* influence the effect *E* independently.
2) *B* could produce *E*, but not prevent it.
3) Causal powers are independent of the frequency of occurrences of the causes.
4) *E* does not occur unless it is caused.

Most of the assumptions have a straightforward interpretation. The major exception is assumption 3, which requires a definition of "causal powers". Cheng (1997, p. 372) defines a causal power as "the probability with which *x* produces *e* when *x* is present," where *x* is some cause and *e* is the effect. On this definition, assumption 3 is difficult to interpret since the occurrence of a cause is an event while causal power is a variable that holds some value on the $[0,1]$ continuum. Thus, the probability that a causal power takes on a point value is zero. A probability distribution for causal powers is necessary to map intervals of values to events. However, elaborating assumption 3 along this tack seems altogether too complicated.

A modified derivation of causal power eliminates the need for assumption 3. To see how, denote c^+ as the event "*C* occurs" and e^+ as the event "*E* occurs". Now introduce new notation of c_e^+ to denote the event "*C* causes *E*". Similarly define the events $b^+$and b_e^+. Clearly, by these definitions c_e^+ is a subset of c^+ and likewise b_e^+ is a subset of b^+.

Hume (1748/1854) famously argued that the connection between cause and effect is beyond perception and that instead we only see the occurrence of causes followed by the occurrence of effects. To respect Hume's argument is to assume that only the c^+ events are observable while the c_e^+ events are not. That said, *C* will be the only plausible cause present in certain contexts. More formally, for some contexts there may be very good reason to believe that $P(b^+) = 0$. In such contexts, one may have high confidence in attributing the occurrence of the effect to the candidate cause. While one may high confidence, they will never be certain. This is because of the standard assumption that the context variable *B* is not directly observable.

From the preceding definitions we have $P(c_e^+|c^+)$ as the probability of the cause producing the effect given that the cause is present, which precisely matches Cheng's (1997) definition of a causal power. The lone, and critical, difference is that I explicitly express causal power as a conditional probability while Cheng (1997) embeds it within her framework as the variable p_x, leading to the confusion regarding assumption 3.

With the new definitions in place, the independence assumption can now be precisely established. Assume independence of the causal generation events c_e^+ and b_e^+. Also assume independence of c^+ and b_e^+, so the occurrence of the candidate cause is independent of the background generation event. The modified primary assumptions then become:

1) c^+ and c_e^+ are both independent of b_e^+
2) *B* could produce *E*, but not prevent it.
3) *E* does not occur unless it is caused.

These assumptions allow the derivation of Cheng's (1997) generative causal power. By assumption 3, the effect occurs only if it is caused by *C* or if it is caused by *B*. In the above notation, this means that the effect occurs only when $c^+ \cap c_e^+$ or $b^+ \cap b_e^+$ is true. Note that $c^+ \cap c_e^+ = c_e^+$ and $b^+ \cap b_e^+ = b_e^+$ since c_e^+ and b_e^+ are subsets. The probability of the union is then:

$$\begin{aligned} P(e^+) &= P(c_e^+ \cup b_e^+) \\ &= P(c_e^+) + P(b_e^+) - P(c_e^+ \cap b_e^+) \end{aligned} \tag{A.1}$$

Then by assumption 1:

$$P(e^+) = P(c_e^+) + P(b_e^+) - P(c_e^+) \times P(b_e^+)$$

We may condition on *C* being absent, which implies $P(c_e^+|c^-) = \frac{P(c^- \cap c_e^+)}{P(c^-)} = \frac{0}{P(c^-)} = 0$. Also, by assumption 1, $P(b_e^+|c^-) = P(b_e^+)$. This gives:

$$P(e^+|c^-) = P(b_e^+) \tag{A.2}$$

With c^- true, only the unobserved background can be present. Since *B* is unobserved, it is not possible to estimate background causal power $P(b_e^+|b^+)$ separate from the probability that *B* occurs, or $P(b^+)$. Consequently, the observed $P(e^+|c^-)$ serves as an estimate for the conjunction probability $P(b^+ \cap b_e^+) = P(b_e^+)$. Though expressed somewhat differently, this result agrees in concept with Cheng's (1997) derivation.

We may similarly condition on *C* being present. This implies $P(c_e^+|c^+) = \frac{P(c^+ \cap c_e^+)}{P(c^+)} = \frac{P(c_e^+)}{P(c^+)}$, so $P(c_e^+|c^+) > P(c_e^+)$ whenever $P(c^+) < 1$. Again by assumption 1, $P(b_e^+|c^+) = P(b_e^+)$, so:

$$P(e^+|c^+) = P(c_e^+|c^+) + P(b_e^+) - P(c_e^+|c^+) \times P(b_e^+) \tag{A.3}$$

Equation (A.3) is just the Noisy-OR parameterization for when the cause is present, but now expressed in the new notation. Substituting (A.2) into (A.3) yields:

$$\begin{aligned} P(c_e^+|c^+) &= \frac{P(e^+|c^+) - P(e^+|c^-)}{1 - P(e^+|c^-)} \\ &= \frac{\Delta P}{1 - P(e^+|c^-)} \end{aligned}$$

And this is just the expression for generative causal power model.

In summary, the derivation introduces the new unobservable events "*C* causes *E"* as c_e^+ and "*B* causes *E*" as b_e^+. This allows for the derivation of causal power with a modified and more precise independence assumption. The result is an expression for the unobservable conditional probability $P(c_e^+|c^+)$ in terms of the observable probabilities $P(e^+|c^+)$ and $P(e^+|c^-)$. Furthermore, the meaning of causal power is clarified by expressing it explicitly as $P(c_e^+|c^+)$.

The above derivation also makes transparent the connection between causal power and the edge weights notation of the noisy-OR parameterization. The edge weight w_1 corresponds to $P(c_e^+|c^+)$, or causal power. And the edge weight w_0 corresponds to $P(b_e^+)$, which in turn equals to the observable probability $P(e^+|c^-)$.

A.2 Relaxing the independence assumption

At various points in this dissertation I consider violations of the independence assumption. Violations may result from a statistical dependence in the occurrences of the causes *C* and *B*. Or it may result from interactions of their causal powers. Equation (A.1) can be used to find a general expression for the model in which independence is not assumed:

$$\begin{aligned} P(e^+) &= P(c_e^+ \cup b_e^+) \\ &= P(c_e^+) + P(b_e^+) - P(c_e^+ \cap b_e^+) \\ &= P(c_e^+) + P(b_e^+) - P(c_e^+|b_e^+) \times P(b_e^+) \end{aligned}$$

Conditioning on the absence of the cause gives:

$$P(e^+|c^-) = P(b_e^+|c^-)$$

Conditioning on the presence of the cause gives:

$$\begin{aligned} P(e^+|c^+) &= P(c_e^+|c^+) + P(b_e^+|c^+) - P(c_e^+|b_e^+, c^+) \times P(b_e^+|c^+) \\ &= P(c_e^+|c^+) + P(b_e^+|c^+) - P(c_e^+|b_e^+) \times P(b_e^+|c^+) \end{aligned}$$

And general expressions are obtained that allow for dependence between *C* and *B*.

The model uncertainty simulation of section 3.12 assumed independence between c^+ and b_e^+, so $P(b_e^+|c^+) = P(b_e^+|c^-) = P(b_e^+)$. The model expressed in weights notation then becomes:

$$\begin{aligned} P(e^+|c^-) &= P(b_e^+) \\ &= w_0 \end{aligned}$$

And,

$$P(e^+|c^+) = w_1 + w_0 - P(c_e^+|b_e^+)\times w_0$$
$$= w_1 + w_0 - w_{1|0}\times w_0$$

where $w_{1|0} = P(c_e^+|b_e^+)$ gives the interaction, but does not refer to any specific edge on the common effect graph. With $0 \leq w_{1|0} < 1$ a large range of models can be represented, though clearly a spectrum of models will also be omitted.

A.3 Preventive causal power

Once again, let *C* be the observed candidate cause, *B* be a collection of unobserved causal factors that are net generative, and *E* be the effect. Now consider the case in which the candidate cause *C* is preventive, meaning that it reduces the probability of *E*. Let the event $c_{\sim e}^+$ denote "*C* prevents *E*". By this definition, $c_{\sim e}^+$ will be contained within c^+. Again, b_e^+ represents the event "*B* causes *E"*. Assumption 1 from above then becomes:

1) c^+ and $c_{\sim e}^+$ are both independent of b_e^+

The effect will occur if it is generated by *B* and not prevented by *C*. This can be expressed formally as the conjunction $b^+ \cap b_e^+$ and $\neg(c^+ \cap c_{\sim e}^+)$. Again, this implies $b^+ \cap b_e^+ = b_e^+$ and $c^+ \cap c_{\sim e}^+ = c_{\sim e}^+$. By assumption 1, the events of the conjunction are independent, so:

$$P(e^+) = P[b_e^+ \cap \neg(c_{\sim e}^+)]$$
$$= P(b_e^+)\times P[\neg(c_{\sim e}^+)]$$
$$= P(b_e^+)\times[1 - P(c_{\sim e}^+)]$$

Conditioning on the presence and absence of C, and using assumption 1, then gives:

$$P(e^+|c^-) = P(b_e^+)$$
$$P(e^+|c^+) = P(b_e^+)\times[1 - P(c_{\sim e}^+|c^+)]$$

Substituting the top equation into the bottom and solving for $P(c_{\sim e}^+|c^+)$ yields preventive power:

$$P(c^{+}_{\sim e}|c^{+}) = \frac{-[P(e^{+}|c^{+}) - P(e^{+}|c^{-})]}{P(e^{+}|c^{-})}$$
$$= \frac{-\Delta P}{P(e^{+}|c^{-})}$$

Note that $P(c^{+}_{\sim e}|c^{+})$ is the probability that *C* prevents *E* conditional on *C* occurring. This generally conforms to Cheng's (1997, p. 375) definition of, "a preventive cause *i* has the power to stop an (otherwise occurring) effect *e* from occurring with probability p_i".

Appendix B.

Mean-squared error of causal power

B.1 Conditional mean-squared error for causal power MLE

In order to find an exact expression for the variance, I assume that w_0 is fixed and known, though below this assumption is relaxed.

Suppose $\widehat{w}_T$ is estimated from a sample of size N with $\widehat{w}_T = \hat{P}(e^+|c^+, b^+) = \frac{N(e^+,c^+)}{N}$. The estimate has a scaled binomial distribution, or $\widehat{w}_T \sim \frac{1}{N}\text{Binom}(w_T, N)$. Note we can also define $\epsilon_T = \widehat{w}_T - w_T$ which allows us to write $\widehat{w}_T = w_T + \epsilon_T$. The error ϵ_T has a scaled binomial distribution shifted by w_T. This implies that $E[\epsilon_T] = E[\widehat{w}_T - w_T] = E[\widehat{w}_T] - w_T = 0$ and $\text{Var}[\epsilon_T] = \text{Var}[\widehat{w}_T] = \frac{w_T(1-w_T)}{N}$.

With fixed w_0 the causal power maximum likelihood estimator is:

$$\widehat{w}_1 = \frac{\widehat{w}_T - w_0}{1 - w_0} \tag{B.1}$$

It is straightforward to show that (B.1) is conditionally unbiased for a given w_1:

$$\begin{aligned} E[\widehat{w}_1|w_1] &= E\left[\frac{\widehat{w}_T - w_0}{1 - w_0}\right] \\ &= E\left[\frac{w_1 + w_0 - w_1 w_0 + \epsilon_T - w_0}{1 - w_0}\right] \\ &= E\left[\frac{w_1(1 - w_0) + \epsilon_T}{1 - w_0}\right] \\ &= w_1 \end{aligned}$$

The second line is from the identity $\widehat{w}_T = w_1 + w_0 - w_1 w_0 + \epsilon_T$. Since the MLE is unbiased, the mean squared error is just the variance. Next, find the conditional MSE with:

$$\begin{aligned}\mathrm{MSE}[\widehat{w}_1|w_1] &= E[(\widehat{w}_1 - w_1)^2] = Var[\widehat{w}_1] \\ &= E\left[\left(\frac{w_1(1-w_0)+\epsilon_T}{1-w_0} - w_1\right)^2\right] \\ &= \left(\frac{1}{1-w_0}\right)^2 E[(\epsilon_T)^2]\end{aligned}$$

Since $E[(\epsilon_T)^2] = \mathrm{Var}[\widehat{w}_T]$, this gives:

$$\mathrm{Var}[\widehat{w}_1] = \left(\frac{1}{1-w_0}\right)^2 \times \frac{w_T(1-w_T)}{N} \tag{B.2}$$

Now expand the expression $w_T = w_1 + w_0 - w_1 \times w_0$ and work through more algebra to get:

$$\mathrm{Var}[\widehat{w}_1] = \frac{(1-w_1)}{N} \times \left[\frac{w_0}{1-w_0} + w_1\right] \tag{B.3}$$

The importance of the base-rate relative to the sample size is made clear from (B.3). As $w_0 \to 1$ the MSE becomes arbitrarily large. The MSE changes in w_0 at the rate of $\frac{\partial}{\partial w_0}\mathrm{Var}[\widehat{w}_1] = \frac{1}{(1-w_0)^2} \times c(N, w_1)$, where $c(N, w_1)$ is a positive constant based on *N* and w_1. Since $0 \leq w_0 \leq 1$, this indicates a quadratic rate of increase in w_0. Similarly, the MSE decreases in *N* at the rate of $\frac{\partial}{\partial N}\mathrm{Var}[\widehat{w}_1] = -\frac{1}{N^2} \times c(w_0, w_1)$, also a quadratic change. Thus, the base rate and sample size have commensurate influence on the mean squared error. Also note that at $w_0 = 1$ the derivative of the MSE is not defined, so the MLE does not exist.

B.2 Mean-squared error for mixed causal power estimator

In fact, (B.3) is not a plausible MSE for causal power. The derivation assumes that generative power is applied regardless of whether $\Delta P > 0$ or $\Delta P < 0$. But a reasonable judge would not apply (B.1) for negative ΔP. The typical assumption is that causal direction is first determined by ΔP, and then generative or preventive power is applied depending on whether it is positive or negative. In amending the above, I will use a simplification: the judge will estimate a causal power of 0 when $\Delta P < 0$. The modified causal power estimator is then:

$$\widehat{w}_1 = \begin{cases} \dfrac{\widehat{w}_T - w_0}{1 - w_0}, & \Delta P \geq 0 \\ 0, & \Delta P < 0 \end{cases}$$

The above estimator will be biased, but will often have a lower MSE. Since $\widehat{w}_T \sim \frac{1}{N} \times \text{Binom}(w_T, N)$, we may find $\Pr(w_0 \leq \widehat{w}_T) = p$ using the $\text{Binomial}(w_T, N)$ distribution. More specifically, the binomial distribution can be used to find $p = \Pr(N \times w_0 \leq N \times \widehat{w}_T)$. The product $N \times \widehat{w}_T$ will always be an integer since $\widehat{w}_T = a/N$ is a sample proportion. However, it is necessary to enforce that $N \times w_0 = c$ is also an integer. In other words, only $w_0 = c/N$ are allowable values for the base rate since non-integer values are not interpretable with respect the binomial distribution.

Thus, $\widehat{w}_1$ has a mixture distribution, taking the value $\frac{\widehat{w}_T - w_0}{1 - w_0}$ with probability p and the value 0 with probability $(1 - p)$. For the two distributions we get the conditional variances $\text{Var}[\widehat{w}_1 | \Delta P \geq 0] = \left(\frac{1}{1 - w_0}\right)^2 \text{Var}[\epsilon_T]$ and $\text{Var}[\widehat{w}_1 | \Delta P < 0] = E[(0 - w_1)^2] = w_1^2$. To find the MSE of a mixture distribution requires the rule of total variance:

$$\text{Var}(X) = E[\text{Var}(X|\Theta)] + \text{Var}[E(X|\Theta)]$$

where Θ tracks the different levels of the mixture distribution. In our problem, $\Theta = 1$ will give the estimator distribution for $\Delta P \geq 0$ and $\Theta = 2$ will give the distribution for $\Theta = 1$. Breaking out the first component:

$$E[\text{Var}(X|\Theta)] = p\left(\frac{1}{1 - w_0}\right)^2 \text{Var}[\epsilon_T] + (1 - p) w_1^2$$

And the second component:

$$\begin{aligned} \text{Var}[E(X|\Theta)] &= E\left[\left(E(X|\Theta)\right)^2\right] - (E[E(X|\Theta)])^2 \\ &= p \times E(X|\Theta = 1)^2 + (1 - p) \times E(X|\Theta = 2)^2 \\ &\quad - \left(p \times E(X|\Theta = 1) + (1 - p) \times E(X|\Theta = 2)\right)^2 \end{aligned}$$

Now $E\left(\widehat{w}_1 \middle| \frac{m}{N} \geq w_0\right) = w_1$ and $E\left[\widehat{w}_1 | \frac{m}{N} < w_0\right] = 0$. Plugging into the above:

$$\begin{aligned}\text{Var}[E(X|\Theta)] &= p\times(w_1)^2 + (1-p)\times 0 - (p\times w_1 + (1-p)\times 0)^2 \\ &= p\times(w_1)^2 - (p\times w_1)^2 \\ &= p\times(w_1)^2\times(1-p)\end{aligned}$$

Putting it together gives the conditional variance:

$$\begin{aligned}\text{Var}[\widehat{w}_1|w_1] &= E[\text{Var}(X|\Theta)] + \text{Var}[E(X|\Theta)] \\ &= p\left(\frac{1}{1-w_0}\right)^2 \text{Var}\left[\epsilon_{Tj}\right] + (1-p)w_1^2 + (1-p)w_1^2\times p \\ &= p\left(\frac{1}{1-w_0}\right)^2 \text{Var}[\epsilon_T] + (1-p)w_1^2(1+p) \\ &= (1-p^2)w_1^2 + p\left(\frac{1}{1-w_0}\right)^2 \text{Var}[\epsilon_T]\end{aligned}$$

The final expression is the well-known $\text{Bias}^2 + \text{Variance}$ result, though now weighted by the mixture probabilities $1-p^2$ and p. Before we saw that $\text{Var}[\epsilon_T]$ has the binomial variance $\frac{w_T(1-w_T)}{N}$. Now ϵ_T has a different distribution since a minimum of $m = w_0\times N$ successes must occur to be in the "generative" portion of the mixture.

Overall, this estimator should have lower MSE than the causal power MLE. However, the variance portion resembles (B.2), and so there will be a similar influence of the base rate w_0 as was found above.

B.3 Taylor approximation of causal power variance

Above w_0 was assumed to be known and fixed, but this will not describe most applications. When the base rate is unknown and random, the causal power MLE is $\widehat{w}_1 = \frac{\widehat{w}_T - \widehat{w}_0}{1-\widehat{w}_0}$. In general, it is not possible to find exact expectations or variances for ratios of random variables. A standard technique is to use a multivariate Taylor expansion to approximate these quantities (e.g. Wolter,

2007). Suppose *X* and *Y* are two random variables, then the Taylor approximation for the ratio of their variance is:

$$\mathrm{Var}\left(\frac{X}{Y}\right) \approx \frac{\mu_X^2}{\mu_Y^2}\left(\frac{\sigma_X^2}{\mu_X^2} - 2\frac{\mathrm{Cov}(X,Y)}{\mu_X\mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2}\right)$$

To see how this applies to the causal power MLE, first note that $\frac{\widehat{w}_T - \widehat{w}_0}{1-\widehat{w}_0} = 1 - \frac{1-\widehat{w}_T}{1-\widehat{w}_0}$. In addition, since $\widehat{w}_T$ and $\widehat{w}_0$ are proportions $\mathrm{Var}(1-\widehat{w}_T) = \mathrm{Var}(\widehat{w}_T)$ and $\mathrm{Var}(1-\widehat{w}_0) = \mathrm{Var}(\widehat{w}_0)$. Finally, by the assumption of statistical independence, $\mathrm{Cov}(\widehat{w}_T, \widehat{w}_0) = 0$. Putting this all together, the approximate variance is then:

$$\begin{aligned}\mathrm{Var}\left(\frac{\widehat{w}_T - \widehat{w}_0}{1-\widehat{w}_0}\right) &\approx \frac{(1-w_T)^2}{(1-w_0)^2}\left(\frac{\sigma_T^2}{(1-w_T)^2} - \frac{\sigma_0^2}{(1-w_0)^2}\right) \\ &= \frac{\sigma_T^2}{(1-w_0)^2} + \frac{(1-w_T)^2}{(1-w_0)^4}\sigma_0^2 \end{aligned} \quad \text{(B.4)}$$

Not surprisingly, a random $\widehat{w}_0$ results in a larger variance. Too see this clearly, another way to express (B.2) above is $\frac{\sigma_T^2}{(1-w_0)^2}$. So it is apparent that the approximate variance (B.4) will be larger whenever $w_0 > 0$.

Of course, w_0 and w_T are unknown and must be estimated with $\widehat{w}_0$ and $\widehat{w}_T$. Using the sample estimates will give an estimated approximate variance:

$$\widehat{\mathrm{Var}}\left(\frac{\widehat{w}_T - \widehat{w}_0}{1-\widehat{w}_0}\right) \approx \frac{\hat{\sigma}_T^2}{(1-\widehat{w}_0)^2} + \frac{(1-\widehat{w}_T)^2}{(1-\widehat{w}_0)^4}\hat{\sigma}_0^2$$

Note that we cannot claim that (B.4) returns the MSE since, with a random denominator, it is no longer obvious that $\frac{\widehat{w}_T - \widehat{w}_0}{1-\widehat{w}_0}$ is unbiased. One could use a Taylor expansion to argue that it is approximately unbiased, which would make (B.4) the "approximate approximate MSE" since it is the approximate variance for the approximate MSE. In fact, exploration with simulated data suggests that (B.4) is a decent approximation of the mean squared error.

While the approximate variance is larger with random $\widehat{w}_0$, our primary interest concerns how reliability changes with changes in the base rate, which cannot be gleaned from simple inspection

of (B.4). Substituting in from (B.3) and the identity $(1 - w_T) = (1 - w_0)(1 - w_1)$ allows for further simplification:

$$\begin{aligned} \mathrm{Var}\left(\frac{\widehat{w}_T - \widehat{w}_0}{1 - \widehat{w}_0}\right) &\approx \frac{(1 - w_1)}{N} \times \left[\frac{w_0}{1 - w_0} + w_1\right] + \frac{w_0(1 - w_T)^2}{(1 - w_0)^3 N} \\ &= \frac{(1 - w_1)}{N}\left[w_1 + \frac{w_0}{1 - w_0}(2 - w_1)\right] \end{aligned} \quad \text{(B.5)}$$

The result is very close to the (B.3) with w_0 fixed. From (B.5) we obtain $\frac{\partial}{\partial w_0}\mathrm{Var}[\widehat{w}_1] = \frac{1}{(1-w_0)^2} \times c(N, w_1)$ and $\frac{\partial}{\partial N}\mathrm{Var}[\widehat{w}_1] = -\frac{1}{N^2} \times c(w_0, w_1)$, so the rate of change is quadratic both in the base rate w_0 and the sample size *N*, which is the same result found above for a fixed w_0. One can also derive a similar result for the mixed causal power estimator from Section B.2. The only difference is that the approximate variance from (B.4) will replace $\mathrm{Var}[\epsilon_T]$. And similar to above, (B.4) cannot be further simplified since σ_T^2 no longer has a binomial distribution.

B.4 Unconditional mean-squared error for causal power MLE

The discussion thus far has conditioned on a particular candidate cause *C* with causal power w_1. Also of interest is average estimator performance as it is applied to multiple different causes. To keep the derivations tractable, assume that the base-rate of the effect w_0 is fixed so that the causal power MLE is given by (B.1) above. The MLE is unconditionally unbiased since:

$$E\big[E[\widehat{w}_1|w_1]\big] = E[w_1] = \theta$$

where θ is the mean of the w_1's. Thus, the unconditional MSE will equal the variance of the causal power MLE. Suppose w_1 has some distribution with mean θ and variance τ^2. Then we can find the unconditional MSE using $\mathrm{MSE}(\widehat{w}_1) = E[\mathrm{MSE}(\widehat{w}_1|w_1)]$. Specifically:

$$
\begin{aligned}
E[\mathrm{Var}(\widehat{w}_1)] &= E\left[\frac{(1-w_1)}{N}\times\left[\frac{w_0}{1-w_0}+w_1\right]\right] \\
&= \frac{1}{N}\times\left[\frac{(1-\theta)w_0}{1-w_0}+\theta-E[w_1^2]\right] \\
&= \frac{1}{N}\times\left[\frac{(1-\theta)w_0}{1-w_0}+\theta-(E[w_1^2]-\theta^2)-\theta^2\right] \\
&= \frac{1}{N}\times\left[\frac{(1-\theta)w_0}{1-w_0}+\theta(1-\theta)-\tau^2\right] \\
&= \frac{(1-\theta)}{N}\left[\frac{w_0}{(1-w_0)}+\theta-\frac{\tau^2}{(1-\theta)}\right]
\end{aligned}
$$

Note that in the derivation above, $\tau^2 = E[w_1^2]-\theta^2$.

When w_1 has a beta(a,b) distribution, then $\tau^2 = \frac{\theta(1-\theta)}{1+\nu}$ where $\nu = a+b$ is equal to the prior sample size. Substituting into the above gives an unconditional MSE of:

$$
\begin{aligned}
E[\mathrm{Var}(\widehat{w}_1)] &= \frac{(1-\theta)}{N}\left[\frac{w_0}{(1-w_0)}+\theta-\frac{\theta}{1+\nu}\right] \\
&= \frac{(1-\theta)}{N}\left[\frac{w_0}{(1-w_0)}+\frac{\nu}{1+\nu}\theta\right]
\end{aligned}
$$

The unconditional MSE is quite similar to the conditional MSE above, as shown in (B.3). Again, the MSE becomes arbitrarily large as $w_0 \to 1$ and the instantaneous change is quadratic in w_0.

Finding the unconditional MSE for the estimator that assigns 0 to $\Delta P < 0$ is a greater challenge. It may not be possible to solve for the expectation of the mixture probability $E[p]$. One would also need to account for the dependence between w_1 and p. A larger w_1 should be associated with a larger p. If these quantities are dependent, then $E[p\times w_1] \neq E[p]\times E[w_1]$.

Appendix C.

Mean-squared error of weighted ΔP

C.1 Conditional mean-squared error for weighted ΔP estimator

C.1.1 Fixed base rate w_0

First consider the weighted ΔP estimator with a fixed base rate so that $\widehat{w}_1 = \widehat{w}_T - (1-\theta)\times w_0$, where θ represents the prior expectation for causal power. Assume $\widehat{w}_T$ is estimated from a sample of size *N*. We can also write $\widehat{w}_T = w_T + \epsilon_T$ where the error has a scaled binomial distribution shifted by w_T (see Appendix B). This implies $E[\epsilon_T] = 0$ and $\mathrm{Var}[\epsilon_T] = \mathrm{Var}[\widehat{w}_T]$. Then the MSE conditional on a particular w_1 is given by:

$$\begin{aligned}
\mathrm{MSE}[\widehat{w}_1|w_1] &= E[(\widehat{w}_1 - w_1)^2] \\
&= E[(w_1 + (\theta - w_1)w_0 + \epsilon_T - w_1)^2] \\
&= E\left[\left(\left((\theta - w_1)w_0\right) + \epsilon_T\right)^2\right] \\
&= \left((\theta - w_1)w_0\right)^2 + E[\epsilon_T^2] \\
&= \left((\theta - w_1)w_0\right)^2 + \frac{w_T(1-w_T)}{N} \qquad \text{(C.1)}
\end{aligned}$$

And this is the familiar $\mathrm{Bias}^2 + \mathrm{Variance}$ formula. Recall that the variance of the causal power MLE continues to increase as $w_0 \to 1$. In contrast, the weighted ΔP MSE is bounded at $(\theta - w_1)^2$ as $w_0 \to 1$. To discover precisely how the MSE changes with the base rate requires expanding the w_T terms to obtain:

$$\mathrm{MSE}[\widehat{w}_1|w_1] = \left((\theta - w_1)w_0\right)^2 + \frac{(1-w_1)(1-w_0) - (1-w_1)^2(1-w_0)^2}{N}$$

So the MSE increases in w_0 at $\frac{\partial MSE}{\partial w_0} = 2(\theta - w_1)w_0 - \frac{2(w_0-1)(w_1-1)^2+(w_1-1)}{N}$, a linear rate of change, compared to the quadratic increase for the causal power MLE. In addition, weighted ΔP also has a quadratic decrease in MSE as sample size *N* increases.

C.1.2 Random base rate $\widehat{w}_0$

Unlike with the causal power MLE, it is straightforward to find the MSE when the base rate is random. Consider the weighted ΔP estimator with both random $\widehat{w}_T$ and $\widehat{w}_0$. The estimator is then $\widehat{w}_1 = \widehat{w}_T - (1-\theta)\widehat{w}_0$. Again, θ represents the prior expectation for causal power. So now, $\widehat{w}_1 = w_1 + w_0 - w_1 w_0 + \epsilon_T - (1-\theta)(w_0 + \epsilon_0)$. The error ϵ_0 is defined analogously to ϵ_T, so it is a scaled binomial shifted by w_0. The conditional MSE is then:

$$\begin{aligned}\mathrm{MSE}[\widehat{w}_1|w_1] &= E[(\widehat{w}_1 - w_1)^2] \\ &= E\left[\left((\theta - w_1)w_0\right)^2 + \epsilon_T^2 + (1-\theta)^2\epsilon_0^2\right] \\ &= \left((\theta - w_1)w_0\right)^2 + \frac{w_T(1-w_T)}{N} + (1-\theta)^2\frac{w_0(1-w_0)}{N}\end{aligned}$$

Again, this is just $\mathrm{Bias}^2 + \mathrm{Variance}$, but now there is one additional term in the variance due to the random base rate $\widehat{w}_0$. As before, the MSE increases linearly and is bounded as $w_0 \to 1$.

C.2 Unconditional mean-squared error for weighted ΔP estimator

To simplify the derivation, again assume w_0 is fixed. Also suppose w_1 has some distribution with mean θ and variance τ^2. The unconditional MSE of the weighted ΔP rule is found with:

$$\begin{aligned}E\left[\mathrm{MSE}[\widehat{w}_1|w_1]\right] &= E\left[\left((\theta - w_1)w_0\right)^2 + \frac{w_T(1-w_T)}{N}\right] \\ &= E\left[\left((\theta - w_1)w_0\right)^2 + \frac{w_T(1-w_T)}{N}\right] \\ &= w_0^2 \times E[(\theta - w_1)^2] + E\left[\frac{w_T(1-w_T)}{N}\right]\end{aligned}$$

By definition we have $E[(\theta - w_1)^2] = \tau^2$. From the MLE derivation we can write:

$$E\left[\frac{w_T(1-w_T)}{N}\right] = (1-w_0)^2 \times \frac{(1-\theta)}{N}\left[\frac{w_0}{(1-w_0)} + \theta - \frac{\tau^2}{(1-\theta)}\right]$$

Plugging into the above gives the unconditional MSE:

$$E\left[\text{MSE}[\widehat{w}_1|w_1]\right] = w_0^2\tau^2 + (1-w_0)^2 \times \frac{(1-\theta)}{N}\left[\frac{w_0}{(1-w_0)} + \theta - \frac{\tau^2}{(1-\theta)}\right]$$

And if we assume that $w_1 \sim \text{Beta}(a, b)$ then this gives:

$$E\left[MSE[\widehat{w}_1|w_1]\right] = w_0^2\tau^2 + \frac{(1-\theta)}{N}\left[w_0(1-w_0) + \frac{\nu}{1+\nu}\theta(1-w_0)^2\right]$$

where $\nu = a + b$. When $w_0 \to 1$ the MSE approaches a maximum of τ^2. Once again, the derivative of the MSE is linear in w_0. Of course, weighted ΔP and the MLE are identical in unconditional MSE when $w_0 = 0$ since they are the same estimator in this case.

Appendix D.

Equilibria of modified Rescorla-Wagner

D.1 Rescorla-Wagner with attenuation parameter κ

Consider an elemental causal induction problem with candidate cause C, background cause B, and effect E. Denote association strengths for the candidate and background causes respectively as V_C and V_B. There are four types of trials $\{(c^+, e^+); (c^+, e^-); (c^-, e^+); (c^-, e^-)\}$ with a, b, c and d respectively giving the frequencies of each type. Suppose $\alpha_0 = \alpha_1$, $\beta_0 = \beta_1$ and a maximum possible association strength of $\lambda = 1$, which are all standard assumptions for the Rescorla-Wagner (R-W) model.

Now introduce an additional parameter $\kappa \in [0,1]$. The κ parameter is incorporated into the R-W model with a "neglect function" $g(C)$ that is given by:

$$g(C) = \begin{cases} \kappa & \text{for } c^+ \text{ trials} \\ 1 & \text{for } c^- \text{ trials} \end{cases}$$

The neglect function will multiply the background strength V_B. The modified R-W model has the following equations for each of the trial types. For the (c^+, e^+) trials:

$$\Delta \mathrm{V_C} = \alpha\beta[1 - (\kappa V_B + V_C)]$$
$$\Delta \mathrm{V_B} = \alpha\beta\ [1 - (\kappa V_B + V_C)]$$

For the (c^+, e^-) trials:

$$\Delta \mathrm{V_C} = \alpha\beta[0 - (\kappa V_B + V_C)]$$
$$\Delta \mathrm{V_B} = \alpha\beta[0 - (\kappa V_B + V_C)]$$

For the (c^-, e^+) trials:

$$\Delta \mathrm{V_B} = \alpha\beta[1 - V_B]$$

And for the (c^-, e^-) trials:

$$\Delta \mathrm{V_B} = \alpha\beta[0 - V_B]$$

So on the c^+ trials the weight given to the background strength V_B is diminished by a factor of κ. Following Chapman & Robbins (1990), it is easy to show that this modified model will give our familiar weighted ΔP rule at equilibrium.

The expected change in V_C is the weighted sum of the two trial types on which it changes, with weights given by the corresponding frequencies from the contingency table. The expectation is:

$$\mathrm{E}[\Delta \mathrm{V_C}] = a{\times}\alpha\beta[1 - (\kappa V_B + V_C)] + b{\times}\alpha\beta[0 - (\kappa V_B + V_C)]$$

Similarly, the average change in V_B equals the sum of the four trial types on which it changes weighted by the frequencies of the different types:

$$\mathrm{E}[\Delta \mathrm{V_B}] = a{\times}\alpha\beta[1 - (\kappa V_B + V_C)] + b{\times}\alpha\beta[0 - (\kappa V_B + V_C)] + c{\times}\alpha\beta[1 - V_B] + d{\times}\alpha\beta[0 - V_B]$$

Since α and β are common to all terms, both equations can be simplified to:

$$\frac{\mathrm{E}[\Delta \mathrm{V_C}]}{\alpha\beta} = a - (a + b)V_C - \kappa(a + b)V_B \qquad \text{(D.1)}$$

and

$$\frac{\mathrm{E}[\Delta \mathrm{V_B}]}{\alpha\beta} = a + c - (a + b)V_C - [\kappa(a + b) + (c + d)]V_B \qquad \text{(D.2)}$$

Learning reaches equilibrium when the expected change of $\mathrm{V_C}$ and $\mathrm{V_B}$ is 0. Setting equation (D.2) equal to 0 and solving for V_B gives:

$$V_B = \frac{a + c - (a + b)V_C}{[\kappa(a + b) + c + d]}$$

Setting (D.1) to 0 and substituting in V_B produces:

$$V_C = \frac{a}{a+b} - \kappa V_B$$

$$= \frac{a}{a+b} - \frac{\kappa[a+c-(a+b)V_C]}{[\kappa(a+b)+c+d]}$$

$$\frac{1}{\kappa}[\kappa(a+b)+c+d]V_C - (a+b)V_C = \frac{1}{\kappa}[\kappa(a+b)+c+d]\frac{a}{a+b} - a - c$$

$$\frac{1}{\kappa}(c+d)V_C = a + \frac{a(c+d)}{\kappa(a+b)} - a - c$$

$$V_C = \frac{a}{a+b} - \kappa \times \frac{c}{c+d}$$

$$V_C = P(e^+|c^+) - \kappa P(e^+|c^-)$$

And we get the familiar weighted ΔP rule. It also follows from the above that the equilibrium background strength is equal to the objective conditional probability, or $V_B = \frac{c}{c+d} = P(e^+|c^-)$.

For the case of preventative power everything proceeds as above except that e^- becomes the focal event, so all trials are "reverse coded". The equations for each of the trial types are then given by:

Trial Type	$\Delta V_C =$	$\Delta V_B =$
(c^+, e^+)	$\alpha\beta[0 - (\kappa V_B + V_C)]$	$\alpha\beta[0 - (\kappa V_B + V_C)]$
(c^+, e^-)	$\alpha\beta[1 - (\kappa V_B + V_C)]$	$\alpha\beta[1 - (\kappa V_B + V_C)]$
(c^-, e^+)	0	$\alpha\beta[0 - V_B]$
(c^-, e^-)	0	$\alpha\beta[1 - V_B]$

Following the above derivation then gives:

$$V_C = \frac{b}{a+b} - \kappa \times \frac{d}{c+d}$$

$$= P(e^-|c^+) - \kappa P(e^-|c^-)$$

And weighted ΔP for preventative power is obtained as the equilibrium. However, the equilibrium background strength is now $V_B = \frac{d}{c+d} = P(e^-|c^-)$.

Importantly, the preventative form mirrors the generative form. Hence, these models predict that judgments will evolve identically for preventative evidence that "mirrors" generative evidence. Generative and preventative evidence mirror one another if one sequence is obtained from the other by changing the e^+'s to the e^-'s. For example, the generative sequence $\{a, a, a, b, c, d, d, d\}$ is mirrored by the preventative sequence $\{b, b, b, a, d, c, c, c\}$.

D.2 Rescorla-Wagner with unequal λ parameters

A weighted ΔP equilibrium can also be achieved if one allows the maximum association strengths (the λ's) to differ across contexts. Denote λ^+ as the maximum association strength for the c^+ context and λ^- as the maximum association strength for the c^- context. Furthermore, assume $\lambda^+ = 1$ and $\lambda^- = \lambda$ with $0 < \lambda < 1$. This modified R-W model has the following equations for each of the trial types:

Trial Type	$\Delta V_C =$	$\Delta V_B =$
(c^+, e^+)	$\alpha\beta[1 - (V_B + V_C)]$	$\alpha\beta[1 - (V_B + V_C)]$
(c^+, e^-)	$\alpha\beta[0 - (V_B + V_C)]$	$\alpha\beta[0 - (V_B + V_C)]$
(c^-, e^+)	0	$\Delta \mathrm{V_B} = \alpha\beta[\lambda - V_B]$
(c^-, e^-)	0	$\Delta \mathrm{V_B} = \alpha\beta[0 - V_B]$

Again, following Chapman & Robbins (1990), the expected change in association strength V_C is:

$$\mathrm{E}[\Delta \mathrm{V_C}] = a{\times}\alpha\beta[1 - (V_B + V_C)] + b{\times}\alpha\beta[0 - (V_B + V_C)]$$

The expected change in V_B is:

$$\mathrm{E}[\Delta \mathrm{V_B}] = a{\times}\alpha\beta[1 - (V_B + V_C)] + b{\times}\alpha\beta[0 - (V_B + V_C)] + c{\times}\alpha\beta[\lambda - V_B] + d{\times}\alpha\beta[0 - V_B]$$

Since α and β are common to all terms, both equations can be simplified by division:

$$\frac{\mathrm{E}[\Delta \mathrm{V_C}]}{\alpha\beta} = a - (a+b)V_C - (a+b)V_B \tag{D.3}$$

and

$$\frac{\mathrm{E}[\Delta \mathrm{V_B}]}{\alpha\beta} = a + \lambda c - (a+b)V_C - [(a+b)+(c+d)]V_B \tag{D.4}$$

Learning reaches equilibrium when the expected change of $\mathrm{V_C}$ and $\mathrm{V_B}$ is 0. Setting equation (D.4) equal to 0 and solving for V_B gives:

$$V_B = \frac{a + \lambda c - (a+b)V_C}{[a+b+c+d]}$$

Setting (D.3) to 0 and substituting in V_B produces:

$$\begin{aligned} V_C &= \frac{a}{a+b} - V_B \\ &= \frac{a}{a+b} - \frac{a + \lambda c - (a+b)V_C}{[a+b+c+d]} \\ \frac{[a+b+c+d]V_C - (a+b)V_C}{[a+b+c+d]} &= \frac{a}{a+b} - \frac{a+\lambda c}{[a+b+c+d]} \\ (c+d)V_C &= a + \frac{a(c+d)}{a+b} - a - \lambda c \\ &= \frac{a}{a+b} - \lambda \times \frac{c}{c+d} \end{aligned}$$

Again we obtain the weighted ΔP rule. Now the weight is given by λ, the maximum association strength for the c^- context. Finally, it follows that the equilibrium association strength given to the background no longer equals the objective conditional probability. Instead, it is now equal to $V_B = \lambda \times \frac{c}{c+d} = \lambda \times P(e^+|c^-)$.

D.3 Modified Rescorla-Wagner converges to causal power

It can be shown that the augmented R-W models will converge to causal power when the additional κ or λ_2 parameters are updated with the association strength. Begin with the κ attenuation model

and set the initial $\kappa = \theta$. After each learning trial the attenuation parameter can be updated with $\kappa = 1 - V_C$. Simply substituting $\kappa = 1 - V_C$ into the above derivation yields the equilibrium strengths:

$$V_C = \frac{a}{a+b} - (1 - V_C) \times \frac{c}{c+d}$$
$$V_C = P(e^+|c^+) - (1 - V_C) \times P(e^+|c^-)$$

Solving for strength V_C then gives causal power at equilibrium:

$$V_C = \frac{P(e^+|c^+) - P(e^+|c^-)}{1 - P(e^+|c^-)}$$

The same argument can be applied to the modified Rescorla-Wagner for preventive power:

$$V_C = \frac{b}{a+b} - (1 - V_C) \times \frac{d}{c+d}$$
$$V_C = P(e^-|c^+) - (1 - V_C) \times P(e^-|c^-)$$

Solving for V_C then yields preventive power at equilibrium:

$$V_C = \frac{P(e^+|c^+) - P(e^+|c^-)}{P(e^+|c^-)}$$

Finally, the exact same arguments can be used to show how updating the λ_2 parameter from the model in (D.2) will also produce an algorithm that converges to causal power.

Danks et al. (2003) also describe a R-W model that converges to power. This is achieved by incorporating the noisy-OR/AND-NOT prediction into an augmented R-W model (this augmented model is described by Van Hamme and Wasserman (1994)). For one candidate generative cause, a simple version of their model can be expressed with the four sets of Δ equations:

Trial Type	$\Delta V_C =$	$\Delta V_B =$
(c^+, e^+)	$\alpha\beta(1 - [1 - (1 - V_C)(1 - V_B)])$	$\alpha\beta(1 - [1 - (1 - V_C)(1 - V_B)])$
(c^+, e^-)	$\alpha\beta(0 - [1 - (1 - V_C)(1 - V_B)])$	$\alpha\beta(0 - [1 - (1 - V_C)(1 - V_B)])$
(c^-, e^+)	0	$\alpha\beta[1 - V_B]$
(c^-, e^-)	0	$\alpha\beta[0 - V_B]$

It is typically assumed that at the first step strengths start at $V_C = 0$ and $V_B = 0$. The product $(1 - V_C)(1 - V_B)$ is the noisy-OR prediction for when the candidate cause is present and V_B is the prediction for when the cause is absent.

The equations in the bottom two rows are the same as those shown for the κ model from Section D.1. Now suppose for the generative κ model that $\theta = 0$ and κ is updated with $(1 - V_1)$ after every learning trial. In this case, the κ model and the noisy-OR prediction model are identical. This is clear by re-arranging the Noisy-OR prediction:

$$[1 - (1 - V_C)(1 - V_B)] = [(1 - V_C)V_B + V_C] = \kappa V_B + V_C$$

For preventative causes, Danks et al. (2003) use a noisy-AND-NOT prediction. A simple version of this model can be expressed with the equations:

Trial Type	$\Delta V_C =$	$\Delta V_B =$
(c^+, e^+)	$\alpha\beta[1 - (1 - V_C)V_B]$	$\alpha\beta[1 - (1 - V_C)V_B]$
(c^+, e^-)	$\alpha\beta[0 - (1 - V_C)V_B]$	$\alpha\beta[0 - (1 - V_C)V_B]$
(c^-, e^+)	0	$\alpha\beta[1 - V_B]$
(c^-, e^-)	0	$\alpha\beta[0 - V_B]$

The connection between the preventative models is more opaque. Part of the complication comes from the fact that V_B estimates $P(e^-|c^-)$ in the κ model from above. Denote $\bar{V}_B = P(e^-|c^-)$, so $\bar{V}_B = 1 - V_B$ and again assume κ is updated on every trial with $1 - V_C$. Then starting with Noisy-AND-NOT:

$$\begin{aligned}(1 - V_C)V_B &= (1 - V_C)(1 - \bar{V}_B) \\ &= 1 - V_C - \bar{V}_B + V_C\bar{V}_B \\ &= 1 - [(1 - V_C)\bar{V}_B + V_C] \\ &= 1 - [\kappa\bar{V}_B + V_C]\end{aligned}$$

Finally, the max association strength $\lambda = 1$ has been switched from e^+ trials to e^- in the κ model due to the reverse coding. So for (c^+, e^+) the predicted change is:

$$\begin{aligned}\Delta V_C &= 0 - (1 - [\kappa\bar{V}_B + V_C]) \\ &= -(1 - [\kappa\bar{V}_B + V_C])\end{aligned}$$

And for (c^+, e^-) :

$$\begin{aligned}\Delta V_C &= 1 - (1 - [\kappa\bar{V}_B + V_C]) \\ &= [\kappa\bar{V}_B + V_C]\end{aligned}$$

Except for the sign change, these are the same predictions as found in Danks et al. (2003) for the corresponding trial types. This suggests that the expression from the Danks et al. (2003) paper may be in error.

Appendix E.

Model uncertainty

In this appendix I evaluate the general parameterization:

$$w_T = w_1 + w_0 - w_{1|0} \times w_0 \tag{E.1}$$

where $w_{1|0} \in [0,1]$ describes the interaction term. Appendix A provides the motivation for this parameterization. Assume that $w_1 = w_{1|0} + \epsilon$ with $E[\epsilon] = 0$ and $\text{Var}[\epsilon] = \sigma^2$ To isolate the influence of model uncertainty, assume that w_T and w_0 are known constants.

E.1 Bias and MSE in the context of model uncertainty

E.1.1 Causal power MLE

The causal power sample estimate is:

$$\begin{aligned}
\widehat{w}_1 &= \frac{w_T - w_0}{1 - w_0} \\
&= \frac{w_1 + w_0 - w_{1|0} w_0 - w_0}{1 - w_0} \\
&= \frac{w_1 - (w_1 + \epsilon) w_0}{1 - w_0} \\
&= w_1 + \frac{w_0}{1 - w_0} \epsilon
\end{aligned}$$

It follows that the causal power estimator is unbiased since $E[\widehat{w}_1] = w_1$. So the MSE is equal to the variance, which is:

$$\mathrm{Var}[\widehat{\mathrm{w}}_1] = \mathrm{Var}\left[\mathrm{w}_1 + \frac{w_0}{1-w_0}\epsilon\right]$$
$$= \left(\frac{w_0}{1-w_0}\right)^2 \sigma^2$$

E.1.2 Weighted ΔP estimator

The weighted ΔP estimator is given by:

$$\widehat{w}_1 = w_1 + (1-\theta)w_0$$

where θ is now a prior expectation for the interaction term $w_{1|0}$. The weighted ΔP bias is:

$$E[(\widehat{w}_1 - w_1)] = E\left[w_T - (1-\theta)w_0 - \left(w_T - \left(1 - w_{1|0}\right)w_0\right)\right]$$
$$= w_0 E\left[\theta\ - w_{1|0}\right]$$
$$= w_0(\theta - w_1)$$

And the MSE is given by:

$$E[(\widehat{w}_1 - w_1)^2] = E\left[\left[w_T - (1-\theta)w_0 - \left(w_T - \left(1 - w_{1|0}\right)w_0\right)\right]^2\right]$$
$$= E\left[\left[\left(\theta - w_{1|0}\right)w_0\right]^2\right]$$
$$= w_0^2 E\left[\left[\theta - w_{1|0}\right]^2\right]$$
$$= w_0^2 E[[\theta - (w_1 + \epsilon)]^2]$$
$$= w_0^2 E[\theta^2 - 2k(w_1 + \epsilon) + (w_1 + \epsilon)^2]$$
$$= w_0^2[(\theta - w_1)^2 + \sigma^2]$$

E.2 Sampling $w_{1|0}$ from a beta distribution

To simulate model uncertainty with $E\left[w_{1|0}\right] = w_1$ we can randomly sample from a $\mathrm{beta}(\alpha, \beta)$ using the mean and "prior sample size" parameterization. To obtain a mean of w_1 simply choose a prior sample size ν and set $\alpha = w_1 \times \nu$ and $\beta = (1 - w_1) \times \nu$. This will give a random draw from

the [0,1] interval for the $w_{1|0}$ interaction parameter. The simulation discussed in the main text uses $\nu = 1$.

One complication is that with $w_1 + w_0 > 1$ the interaction parameter $w_{1|0}$ must have a nonzero minimum so that w_T is a legal probability. With the (E.1) parameterization and $w_1 + w_0 > 1$ we must have $w_{1|0} \geq \frac{(w_1+w_0)-1}{w_0}$. In this case, sampling from a $\text{beta}(\alpha, \beta)$ will require a couple of steps. Begin by defining $c = \frac{(w_1+w_0)-1}{w_0}$ and $\mu_Z = \frac{w_1-c}{1-c}$. Expanding μ_Z gives:

$$\begin{aligned}\mu_Z &= \frac{w_1 - c}{1 - c} \\ &= \frac{w_1 - \frac{(w_1 + w_0) - 1}{w_0}}{1 - \frac{(w_1 + w_0) - 1}{w_0}} \\ &= 1 - w_0\end{aligned}$$

And so $0 \leq \mu_Z \leq 1$. Thus, μ_Z is an appropriate mean for a beta distribution.

Hence, when $w_1 + w_0 > 1$, the first step is to sample Z from a $\text{beta}(\alpha, \beta)$ distribution with $E[Z] = \frac{w_1-c}{1-c} = \mu_Z$. This can be done as before by setting $\alpha = \mu_Z \times \nu$ and $\beta = (1 - \mu_Z) \times \nu$. Then Z will be distributed on the interval [0,1]. We may also find the variance $\sigma^2 = \frac{\mu_Z \times (1-\mu_Z)}{(1+\nu)}$, which is a general result concerning the beta distribution.

With Z sampled from this beta distribution, take the linear transformation $X = c + (1 - c)Z$. Since $Z \in [0,1]$ it is easy to see that $X \in [c, 1]$ and we have the appropriate minimum and maximum. Furthermore, since it is a linear transformation we know that X also has a beta distribution. Finally, we may find the mean and variance of X.

$$\begin{aligned}E[X] &= E[c + (1 - c)Z] \\ &= c + (1 - c)E[Z] \\ &= c + (1 - c)\frac{w_1 - c}{1 - c} = w_1\end{aligned}$$

The expectation is again at w_1, giving causal power as the "average" model. And the variance is

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{Var}[c + (1-c)Z] \\ &= (1-c)^2 \times \sigma^2\end{aligned}$$

So the variance is reduced as the minimum value *c* increases.

Appendix F.

Deterministic bias and Bayesian estimation

In this appendix I show that no Bayesian model of parameter estimation can give deterministic predictions of 0 for the [0,8,0,8] condition, 1 for the [8,0,0,8] condition and probabilistic ratings for the remaining conditions.

Let *G* be a common effect causal graph that has an always present background cause *B* with edge weight w_0 and a binary candidate cause *C* with edge weight w_1. Further, suppose that the weights combine according to the noisy-OR function to give the probability that a binary effect *E* will occur.

The learning data $D = [a, b, c, d]$ gives entries from a 2x2 contingency table where *a* gives the frequency of (e^+, c^+) outcomes, *b* the (e^-, c^+) outcomes, *c* the (e^+, c^-) outcomes and *d* the (e^-, c^-) outcomes. In the discussion below, assume that $a, b, c, d > 0$ unless explicitly stated otherwise.

Suppose a judge's prior beliefs over the hypothesis space of edge weights is described by the joint probability distribution $f(w_0, w_1)$. For a given set of learning data, Bayesian inference combines the prior and the noisy-OR likelihood to form a posterior distribution $f(w_0, w_1|D)$. The Bayesian point estimate of causal strength will be given by the posterior expectation for w_1, as this is the standard choice in the literature for models of causal judgment.

Theorem. Consider the data vectors $[0, b, 0, d]$ and $[a, 0, 0, d]$ where $a, b, d > 0$. Then no Bayesian model of parameter estimation for graph *G* can return posterior expectations $E(w_1|D = [0, b, 0, d]) = 0$, $E(w_1|D = [a, 0, 0, d]) = 1$ and also $0 < E(w_1|D') < 1$ for $D' \notin \{[0, b, 0, d];\ [a, 0, 0, d]\}$, where data vectors in D' have at least two observations, one from each row of the contingency table.

Proof. First I show that Bayesian inference over data vectors $[0, b, 0, d]$ and $[a, 0, 0, d]$ will return posterior expectations of $E[w_1|D] = 0$ and $E[w_1|D] = 1$ if and only if the prior distribution $f(w_0, w_1)$ has the form:

$$f(w_0, w_1) = \begin{cases} \alpha_1 \quad \text{for } (0,0) \\ \alpha_2 \quad \text{for } (0,1) \\ \alpha_3 \text{ for } (1, w_1') \\ 0 \quad \text{otherwise} \end{cases} \tag{F.1}$$

where $w_1' \in [0,1]$ and $\alpha_1, \alpha_2 > 0$ and $\alpha_3 \geq 0$. In other words, there must be some positive prior probability allocated to the two pairs of points $(0,0)$ and $(0,1)$. And the only additional positive prior probability may be allocated to the line $(1, w_1')$ with $w_1' \in [0,1]$.

I begin by finding the class of priors that give $E[w_1|D] = 1$ for data vectors of the form $[a, 0,0, d]$. Suppose that the prior $f(w_0, w_1)$ only has support on $\{(0,0); (0,1); (1, w_1')\}$, with the vector $(\alpha_1, \alpha_2, \alpha_3)$ denoting the probability allocated to each element. Then for the data $[a, 0,0, d]$, the posterior expectation is $E[w_1|D] = 1$ if and only if $\alpha_1, \alpha_3 \geq 0$ and $\alpha_2 > 0$.

Proof **(→).** Assume $\alpha_1, \alpha_3 \geq 0$ and $\alpha_2 > 0$ for the prior $f(w_0, w_1)$ over $\{(0,0); (0,1); (1, w_1')\}$. The likelihoods for each of the three regions of positive support are:

$$f(D|w_0 = 0, w_1 = 0) = (1)^d (0)^a = 0$$
$$f(D|w_0 = 0, w_1 = 1) = (1)^d (1)^a = 1$$
and
$$f(D|w_0 = 1, w_1 = w_1') = (0)^d (w_1')^a = 0, \text{ for } \forall\, w_1' \in [0,1]$$

The probability of the data $\Pr(D)$ is then:

$$\Pr(D) = \sum_{(w_0, w_1)} f(D, w_0, w_1) = \sum_{(w_0, w_1)} f(D|w_0, w_1)\, f(w_0, w_1)$$
$$= 0 \times \alpha_1 + 1 \times \alpha_2 + 0 \times \alpha_3 = \alpha_2$$

And the posterior distribution is found with Bayes rule:

$$f(w_0 = 0, w_1 = 0|D) = \frac{0 \times \alpha_1}{\alpha_2} = 0$$
$$f(w_0 = 0, w_1 = 1|D) = \frac{1 \times \alpha_2}{\alpha_2} = 1$$
$$f(w_0 = 1, w_1 = w_1'|D) = \frac{(0)^d (w_1')^a \times \alpha_3}{\alpha_2} = 0$$

Then posterior expectation is:

$$E[w_1|D] = \sum_{(w_0,w_1)} \int w_1 \times f(w_0, w_1|D)$$

$$= 0\times 0 + 1\times 1 + \int_0^1 (0\times w_1) dw_1' = 1$$

And so $E[w_1|D] = 1$ as desired.

(←). Now suppose to the contrary that $E[w_1|D] = 1$ but that $f(w_0, w_1)$ has positive support outside of $\{(0,0); (0,1); (1, w_1)\}$. In particular, for the point (w_0', w_1') suppose positive support $f(w_0', w_1') = \alpha_4 > 0$, where $0 < w_0', w_1' < 1$. Again, suppose $f(0,1) = \alpha_2$. Then the likelihoods are given by:

$$f(D|w_0 = w_0', w_1 = w_1') = (w_0')^d (w_1' + w_0' - w_1' w_0')^a = y$$

and

$$f(D|w_0 = 0, w_1 = 1) = (1)^d (1)^a = 1$$

The other two likelihoods are omitted, as they will again be zero. Since $0 < w_0', w_1' < 1$, the above implies that $0 < y < 1$. Then the probability of the data is:

$$\Pr(D) = y\times \alpha_4 + 1\times \alpha_2$$

The posterior distribution is then:

$$f(w_0 = w_0', w_1 = w_1'|D) = \frac{y\alpha_4}{y\alpha_4 + \alpha_2}$$

and

$$f(w_0 = 0, w_1 = 1|D) = \frac{\alpha_2}{y\alpha_4 + \alpha_2}$$

Let $\beta = \frac{y\alpha_4}{y\alpha_4 + \alpha_2}$, then $(1 - \beta) = \frac{\alpha_2}{y\alpha_4 + \alpha_2}$. The posterior expectation is:

$$E[w_1|D] = w_1' \times \beta + 1\times (1 - \beta)$$

And so $E[w_1|D] < 1$, a contradiction. Thus, it has been shown that $E[w_1|D] = 1$ if and only if the prior $f(w_0, w_1)$ has support on $\{(0,0); (0,1); (1, w_1')\}$ with $f(0,1) > 0$.

Now we also want the prior to give posterior expectation $E[w_1|D] = 0$ for the data vectors of the form $[0, b, 0, d]$. Can this be achieved using the family of priors identified in (F.1) above? Once again, suppose that the prior $f(w_0, w_1)$ only has support on $\{(0,0); (0,1); (1, w_1')\}$, with the vector $(\alpha_1, \alpha_2, \alpha_3)$ denoting the probability allocated to each element. Then for the data $[0, b, 0, d]$, the posterior expectation is $E[w_1|D] = 0$ if and only if $\alpha_1 > 0$ and $\alpha_2, \alpha_3 \geq 0$.

Proof. The proof is identical to the one above. The only difference is that now $f(D|w_0 = 0, w_1 = 0) = 1$ while the other likelihoods equal zero. The upshot is that there must be some positive prior probability on $(0,0)$ in order to obtain $E[w_1|D] = 0$.

In summary, for a noisy-OR likelihood and prior $f(w_0, w_1)$, Bayesian updating will give posterior expectations $E(w_1|D = [0, b, 0, d]) = 0$ and $E(w_1|D = [a, 0,0, d]) = 1$ so long as some positive prior probability is given to the hypotheses $(0,0)$ and $(0,1)$. Positive prior probability may also be allocated to $(1, w_1')$ with $w_1' \in [0,1]$, though it is not necessary. The prior may not have support beyond these two points and this line. We can now use this fact to prove the initial theorem. This is achieved by examining the Bayesian predictions obtained from applying this prior to the remaining patterns of learning data.

Proof. Above we showed that the prior $f(w_0, w_1)$ must have positive support only on $\{(0,0); (0,1); (1, w_1')\}$ where some positive probability is required for the first two elements while it is optional for the third element. Now we are interested in applying this prior to data vectors D' with the form $\{[0, b, c, 0]; [0, b, c, d]; [a, 0, c, d]; [a, b, 0, d]; [a, b, c, 0]; [a, b, c, d]\}$. Each of these vectors receives no prior probability from $f(w_0, w_1)$ and so for each $\Pr(D) = 0$. For example, with the data $[0, b, c, d]$ we get:

$$f(D|w_0 = 0, w_1 = 0) = (1)^d(0)^a(1)^b = 0$$
$$f(D|w_0 = 0, w_1 = 1) = (1)^d(1)^a(0)^b = 0$$
and
$$f(D|w_0 = 1, w_1 = w_1') = (0)^d(w_1')^a(1 - w_1')^b = 0$$

With $\Pr(D) = 0$, the model gives no posterior prediction. This is because the data completely contradict prior beliefs. In summary, the family of priors that give deterministic predictions of 0 for the [0,8,0,8] condition and 1 for the [8,0,0,8] condition will not give any predictions for the remaining conditions.

***Corollary*.** The above theorem holds for any generating function that has:

For $D = [a, 0,0, d]$	For $D = [0, b, 0, d]$
$f(D\|w_0 = 0, w_1 = 0) = 0$	$f(D\|w_0 = 0, w_1 = 0) = 1$
$f(D\|w_0 = 0, w_1 = 1) = 1$	$f(D\|w_0 = 0, w_1 = 1) = 0$
$f(D\|w_0 = 1, w_1 = w_1') = 0$	$f(D\|w_0 = 1, w_1 = w_1') = 0$
$f(D\|w_0 = w_0', w_1 = w_1') = y$	$f(D\|w_0 = w_0', w_1 = w_1') = y$

where $0 < w_0', w_1' < 1$ and $0 < y < 1$. The corollary is easily verified using the above proofs. It would seem that most any reasonable generating function will satisfy the requirements of the corollary.

In sum, the above findings have serious implications for Bayesian models of strength estimation. Most people give ratings of 0 for data $[0, b, 0, d]$, ratings of 1 for data $[a, 0,0, d]$, and probabilistic ratings for the remaining contingency tables. The above theorem implies that this pattern of judgment is strongly inconsistent with any Bayesian model of causal power. And the corollary extends the result to an extremely broad class of generating functions.

Appendix G.

Two-stage causal inference

G.1 Posterior probabilities for model selection

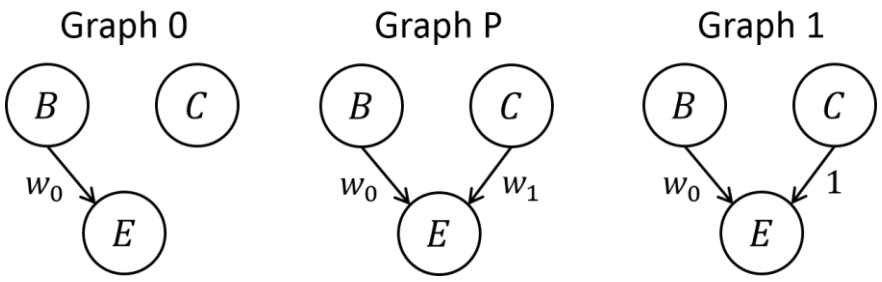


Figure G.1. Directed graphs with B representing the background variable, C the candidate cause, and E the effect of interest. Graph 0 represents a deterministic hypothesis of no causal strength, or $w_1 = 0$. Graph P represents the hypothesis of probabilistic causal strength with $0 < w_1 < 1$. And Graph 1 represents the deterministic hypothesis of $w_1 = 1$.

Consider the three models represented by the three graphs in Figure G.1. Note that Graph 0 implies $w_1 = 0$, Graph P allows $0 < w_1 < 1$ and Graph 1 implies $w_1 = 1$. The posterior probability for each graph is found using Bayes rule:

$$\Pr(\text{Graph i}|D) = \frac{\Pr(D|\text{Graph i}) \times \Pr(\text{Graph i})}{\Pr(D)}$$

Assume equal prior weight is given to each graph so $\Pr(\text{Graph 0}) = \Pr(\text{Graph P}) = \Pr(\text{Graph 1}) = \frac{1}{3}$. To apply Bayes rule we also need to compute $\Pr(\text{Graph i}|D)$ for each graph. This can be achieved using the approach from the appendix of Griffiths and Tenenbaum (2005). Let $D = [a, b, c, d]$ represent the entries of the contingency table. Then the probability of Graph 0 is found with:

$$\Pr(D|\text{Graph } 0) = \int_0^1 p(D|w_0, \text{Graph } 0)\, p(w_0|\text{Graph } 0) dw_0$$

The $p(D|w_0, \text{Graph } 0)$ is just a Bernoulli likelihood with probability w_0 while $p(w_0|\text{Graph } 0)$ is the prior distribution for w_0 on Graph 0. Assume that $p(w_0|\text{Graph } 0)$ is a uniform prior, then:

$$\begin{aligned}\Pr(D|\text{Graph } 0) &= \int_0^1 w_0^{(a+c)} \times (1-w_0)^{(b+d)}\, dw_0 \\ &= \text{B}[a+c+1, b+d+1]\end{aligned}$$

where $\text{B}(x, y)$ is the beta function.

The derivation for $\Pr(D|\text{Graph } 1)$ is similar. Once again assume a uniform prior for $p(w_0|\text{Graph } 1)$. Now $w_1 = 1$ in the likelihood gives:

$$\begin{aligned}\Pr(D|\text{Graph } 1) &= \int_0^1 (1)^a \times (0)^b \times w_0^c \times (1-w_0)^d\, dw_0 \\ &= \begin{cases} 0 & \text{for } b > 0 \\ \text{B}[c+1, d+1] & \text{for } b = 0 \end{cases}\end{aligned}$$

Finding $\Pr(D|\text{Graph P})$ is the most complex with:

$$\Pr(D|\text{Graph P}) = \int_0^1 \int_0^1 p(D|\text{Graph P}, w_0, w_1) \times p(w_0, w_1|\text{Graph P}) dw_0 dw_1 \qquad \text{(G.1)}$$

where $p(D|\text{Graph P}, w_0, w_1)$ is a Bernoulli sampling model with probability given by the Noisy-OR likelihood. Also assume that $p(w_0, w_1|\text{Graph P})$ is a joint uniform prior. The integral (G.1) cannot be solved analytically, so must be approximated.

Since the support is only over a unit-square, the integral can be approximated using a fine grid. One can divide the unit square into *m* equally-sized smaller squares, so each smaller square has area $1/m$. A rectangular box approximates the volume under the i^{th} square. Specifically, one can sample a point (w_{0i}, w_{1i}) within the square and find:

$$V(w_{0i}, w_{1i}) \approx \frac{1}{m} \times p(D|G, w_{0i}, w_{1i}) \times p(w_{0i}, w_{1i}|\text{Graph P})$$

where $1/m$ gives the area of the base and $p(D|G, w_{0i}, w_{1i})\times p(w_{0i}, w_{1i}|\text{Graph P})$ gives the height. And so the approximate integral is found by summing over all of the volumes:

$$\Pr(D|\text{Graph P}) \approx \sum_{i=1}^{m} \frac{1}{m} \times p(D|G, w_{0i}, w_{1i}) \times p(w_{0i}, w_{1i}|\text{Graph P})$$

This is essentially the same approach as that of Griffiths and Tenenbaum (2005), though instead of forming a grid they randomly sample points within the unit square.

G.2 Two-stage Bayesian model to mimic weighted ΔP

Begin with the three graphs from above and assume the Noisy-OR parameterization for Graph P. To mimic weighted ΔP predictions, Graph 0 should be selected for the [0,8,0,8] condition, Graph 1 for the [8,0,8,0] condition, and Graph P should be selected for all remaining conditions. In order for these selections to occur, the prior weight given to each graph had to be carefully chosen. With prior probabilities $P(\text{Graph 0}) = P(\text{Graph 1}) = 1/7$ and $P(\text{Graph P}) = 5/7$ the correct model was generally chosen, though not always.

In addition, the prior on Graph P must be constructed to give estimates close to the weighted ΔP model. In particular, for conditions with observed $P(e^+|c^-) = 0$, the Bayesian prediction should approximately equal the observed $P(e^+|c^+)$. This may be achieved with the prior:

$$\begin{aligned} p(w_0, w_1) &= \text{Beta}[w_0, \alpha_0, \beta_0]\times\text{Beta}[w_1, \epsilon + (w_0)^k\times(\alpha_1 - \epsilon), \epsilon + (w_0)^k\times(\beta_1 - \epsilon)] \\ &\propto w_0^{\alpha_0 - 1}(1 - w_0)^{\beta_0 - 1}\times w_1^{[\epsilon + (w_0)^k\times(\alpha_1 - \epsilon) - 1]}(1 - w_1)^{[\epsilon + (w_0)^k\times(\beta_1 - \epsilon) - 1]} \end{aligned}$$

where $0 < \epsilon, k < 1$ and ϵ is chosen to be "small". Setting $\alpha_0 = \beta_0 = \alpha_1 = \beta_1 = 1$ then gives:

$$p(w_0, w_1) \propto w_1^{[\epsilon + (w_0)^k\times(1 - \epsilon) - 1]}(1 - w_1)^{[\epsilon + (w_0)^k\times(1 - \epsilon) - 1]} \qquad \text{(G.2)}$$

So the prior is the product of two distributions with a dependence between w_0 and w_1. As the $w_0 \to 0$ the prior for w_1 approaches a $\text{Beta}[\epsilon, \epsilon]$. This returns a posterior expectation that is close to the sample estimate for causal power, which is also the weighted ΔP prediction. With $w_0 \to 1$ the prior approaches joint uniform. And recall that weighted ΔP and the uniform prior model were generally close in their predictions for $w_0 > 0$.

Using (G.2) for the prior distribution and a noisy-OR likelihood yields a posterior distribution for which there is no analytic solution. The posterior must be approximated using the procedure described in Section G.1. The results presented in Figure 4.6 are from a model that used $\epsilon = 0.1$ and $k = 0.5$.

Appendix H.

Latent variable results

H.1 Exponential density gives causal power predictions

Suppose the response function $F(.)$ follows a cumulative exponential density. Then the response curve, response and inverse response functions are given by:

$$f(\alpha) = \begin{cases} \lambda e^{-\lambda\alpha} & \text{for } 0 \leq \alpha \\ 0 & \text{for } \alpha < 0 \end{cases}$$

$$F(\alpha) = \begin{cases} 1 - e^{-\lambda\alpha} & \text{for } \alpha \geq 0 \\ 0 & \text{for } \alpha < 0 \end{cases} \qquad F^{-1}(w) = -\tfrac{1}{\lambda} \times \log(1 - \mathrm{w}) \qquad 0 \leq \mathrm{w} < 1$$

where $\lambda > 0$ is the rate parameter. For this family of functions, the 3 steps of the latent variable procedure from Section 5.2.2 will give the predictions of the causal power model. Recall steps 1 is to infer the capacities associated with *B* and *B&C*. Step 2 is to take their difference to find the capacity of *C* alone:

$$\begin{aligned} \alpha_1 &= F^{-1}(w_T) - F^{-1}(w_0) \\ &= \left(\frac{1}{\lambda}\right) \times [-\log(1 - w_T) + \log(1 - w_0)] \end{aligned}$$

Step 3 then finds the causal strength of the candidate cause *C*:

$$\begin{aligned} F(\alpha_1) &= 1 - e^{-\lambda \times \left(\frac{1}{\lambda}\right)[-\log(1-w_T)+\log(1-w_0)]} \\ &= 1 - e^{\log(1-w_T)} \times e^{-\log(1-w_0)} \\ &= 1 - e^{\log(1-w_T)} \times e^{\log\left[\frac{1}{(1-w_0)}\right]} \\ &= 1 - \frac{(1 - w_T)}{(1 - w_0)} \\ &= \frac{w_T - w_0}{1 - w_0} \end{aligned}$$

Hence, any exponential distribution will return predictions of the causal power model, with λ simply serving as a scaling parameter.

H.2 Relationship between ΔP and causal power response functions

Next I show that ΔP and causal power constitute points along a continuum of models within the latent variable framework. Recall that the ΔP rule may be represented as a CARP model using a beta(1,1) response curve. In addition, causal power can be represented with any exponential density as the response curve. To show that these models lie along a continuum within the CARP framework is to show that there are a continuum of response curves between the two models, so that one model can be gradually transformed into the other. Specifically, I demonstrate that as one moves from a beta(1,1) to a beta(1, *n*) density, the response curve converges to an exponential distribution as *n* becomes large.

One property of the beta distribution is that $\lim_{n\to\infty} n\times\text{beta}(1,n) \rightarrow \exp(\lambda = 1)$. Thus, if we use $n\times\text{beta}(1,n)$ as the response function, predictions will approach causal power as *n* becomes large. With regards to CARP predictions, multiplying the beta distribution by *n* will only serve to scale the latent capacity dimension. Accordingly, it is not hard to imagine why predictions from a $\text{beta}(1,n)$ response function will also converge to causal power as *n* becomes large. This can be explicitly demonstrated by writing out CARP predictions using the $n\times\text{beta}(1,n)$ distribution and then showing that they are identical to predictions given by a $\text{beta}(1,n)$ response function.

Begin with the $n\times\text{beta}(1,n)$ distribution. The response and inverse response functions are:

$$\begin{aligned}\Pr(nX \leq x) &= \Pr\left(X \leq \frac{x}{n}\right) \\ &= F\left(\frac{x}{n}\right) = 1 - \left(1 - \frac{x}{n}\right)^n \\ F^{-1}(y) &= n - n\sqrt[n]{1-y}\end{aligned}$$

From above, one can see that $F\left(\frac{x}{n}\right)$ goes to $1 - e^{-x}$ as $n \rightarrow \infty$ since by definition of the exponential function, $e^{-x} = \lim_{n\to\infty}\left(1 - \frac{x}{n}\right)^n$. For a given w_0 and w_T we may apply the CARP procedure to find:

$$\begin{aligned}\alpha_1 &= F^{-1}(w_T) - F^{-1}(w_0)\\ &= \left(n - n\sqrt[n]{1-w_T}\right) - \left(n - n\sqrt[n]{1-w_0}\right)\\ &= n\sqrt[n]{1-w_0} - n\sqrt[n]{1-w_T}\\ F(\alpha_1) &= 1 - \left(1 - \frac{n\sqrt[n]{1-w_0} - n\sqrt[n]{1-w_T}}{n}\right)\\ &= 1 - \left(1 - \left[\sqrt[n]{(1-w_0)} - \sqrt[n]{(1-w_T)}\right]\right)^n\end{aligned} \tag{H.1}$$

And so the causal strength prediction is given by (H.1).

We may also find the expression using a beta(1, n) distribution, which has the following response curve, response function and inverse response function:

$$f(\alpha) = \begin{cases} n(1-\alpha)^{n-1} & \text{for } 0 \le \alpha \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$F(\alpha) = \begin{cases} 0 & \text{for } \alpha < 0 \\ 1-(1-\alpha)^n & \text{for } 0 \le \alpha \le 1 \\ 1 & \text{for } \alpha > 1 \end{cases} \qquad F^{-1}(w) = \begin{cases} 0 & \text{for } w < 0 \\ 1-\sqrt[n]{(1-w)} & \text{for } 0 \le w \le 1 \\ 1 & \text{for } w > 1 \end{cases}$$

So for a given w_T and w_0, causal strength is found with:

$$\begin{aligned}\alpha_1 &= F^{-1}(w_T) - F^{-1}(w_0)\\ &= 1 - \sqrt[n]{(1-w_T)} - \left(1 - \sqrt[n]{(1-w_0)}\right)\\ &= \sqrt[n]{(1-w_0)} - \sqrt[n]{(1-w_T)}\end{aligned}$$

And plugging into $F(.)$ gives:

$$F(\alpha_1) = 1 - \left(1 - \left[\sqrt[n]{(1-w_0)} - \sqrt[n]{(1-w_T)}\right]\right)^n \tag{H.2}$$

Since (H.1) and (H.2) are identical, it is evident that the $n\times\text{beta}(1,n)$ and $\text{beta}(1,n)$ distributions give identical predictions.

In summary, it has been shown that ΔP and causal power exists along a continuum of CARP models since a $\text{beta}(1,1)$ response function gives ΔP predictions while $\lim_{n\to\infty} \text{beta}(1,n)$ converges to causal power predictions.

H.3 Weighted ΔP fails as a normative model

The weighted ΔP rule suffers from serious deficiencies as a description of a causal system. Consider the 1-parameter weighted ΔP model:

$$w_1 = w_T - k \times w_0$$

where $k \in [0,1]$. For the weighted ΔP model, it can be shown that no response function $F(.)$ with continuous response curve $f(.)$ exists unless $k = 1$.

Proof. Suppose there exists a response function $F(.)$ with continuous derivative $f(.)$ over its domain. Also suppose a candidate cause C with associated capacity α_1. Evaluate the cause with respect to two different contexts B and B' with associated capacities $\alpha_0 < \alpha_0'$. Then we have:

$$w_0 = F(\alpha_0); \quad w_0' = F(\alpha_0')$$
$$w_T = F(\alpha_0 + \alpha_1); \quad w_T' = F(\alpha_0' + \alpha_1)$$

Causal strength by the weighted ΔP rule is equal to:

$$w_1 = w_T - k w_0 = F(\alpha_0 + \alpha_1) - k \times F(\alpha_0)$$
$$w_1 = w_T' - k w_0' = F(\alpha_0' + \alpha_1) - k \times F(\alpha_0')$$

Now if the rule is consistent, then the same w_1 should be recovered regardless of context. This implies:

$$F(\alpha_0' + \alpha_1) - k \times F(\alpha_0') = F(\alpha_0 + \alpha_1) - k \times F(\alpha_0)$$
$$F(\alpha_0' + \alpha_1) - F(\alpha_0 + \alpha_1) = k \times [F(\alpha_0') - F(\alpha_0)] \qquad \text{(H.3)}$$

Begin with the right side of (H.3) and let $\alpha_0' = \alpha_0 + \epsilon$. Then:

$$\frac{F(\alpha_0 + \epsilon) - F(\alpha_0)}{(\alpha_0 + \epsilon) - \alpha_0} = \frac{F(\alpha_0 + \epsilon) - F(\alpha_0)}{\epsilon}$$
$$\lim_{\epsilon \to 0} \left(\frac{F(\alpha_0 + \epsilon) - F(\alpha_0)}{\epsilon} \right) = F'(\alpha_0) = f(\alpha_0)$$

Similarly for the right side of (H.3),

$$\lim_{\epsilon \to 0}\left(\frac{F(\alpha_0 + \epsilon + \alpha_1) - F(\alpha_0 + \alpha_1)}{\epsilon}\right) = F'(\alpha_0 + \alpha_1) = f(\alpha_0 + \alpha_1)$$

Thus, we may divide both sides of (H.3) by ϵ and take the limit for:

$$\lim_{\epsilon \to 0}\left[\frac{F(\alpha_0' + \alpha_1) - F(\alpha_0 + \alpha_1)}{\epsilon} = k \times \frac{F(\alpha_0') - F(\alpha_0)}{\epsilon}\right]$$

$$f(\alpha_0 + \alpha_1) = k \times f(\alpha_0)$$

Recall that we assumed a continuous response curve function $f(.)$. So the final equality will only hold in general for:

$$\lim_{\alpha_1 \to 0}[f(\alpha_0 + \alpha_1) = k \times f(\alpha_0)]$$

$$f(\alpha_0) = k \times f(\alpha_0)$$

But this implies $k = 1$, which in turn implies the ΔP model with a uniform response curve, or $f(\alpha) = 1$ for all α. Thus, the only allowable weight for a continuous response curve gives the standard ΔP rule.

The above shows an example application of the latent variable framework. However, in this case it is not necessary to show that weighted ΔP is generally non normative. For instance, suppose two different causes C_1 and C_2 and a context B with no causal strength:

$$P(e^+|c_1^-, c_2^-, b^+) = w_0 = 0; \quad P(e^+|c_1^+, c_2^-, b^+) = w_1; \quad P(e^+|c_1^-, c_2^+, b^+) = w_2$$

The problem is that the total probability is not commutative on the conjoining of causes. Suppose cause C_2 is introduced to the context $(C_1 \cap B)$. Then by weighted ΔP,

$$w_2 = w_T - k \times w_1$$
$$w_T = w_2 + k \times w_1$$

Similarly, suppose cause C_1 is introduced to the context $(C_2 \cap B)$. This gives:

$$w_T = w_1 + k \times w_2$$

Putting them together:

$$w_2 - w_1 = k(w_2 - w_1)$$

And this equality will not hold generally unless $k = 1$, again giving the ΔP rule. It will also hold in the special case of $w_2 = w_1$.

In summary, it has been shown that for a weight $k < 1$ the weighted ΔP rule is generally not consistent, and so it fails as a potential normative model.

Appendix I.

Empirical response functions

I.1 Simple algorithm for estimating response functions

Suppose we want to estimate the unknown response function $F(\alpha|\theta_1, \dots, \theta_k)$. At present, assume the response function comes from a single family of densities, but that the particular density is unknown. For example, $F(x|\mu, \sigma^2)$ represents the single family of normal densities while a particular density is given by specific values of μ and σ^2.

We are also interested in the set of causes $\{B, C_1, \dots C_n\}$ with unknown capacities $[\alpha_0, \alpha_1, \dots, \alpha_n]$. As before, B represents the always present context. The causal strength of cause C_i is defined as the probability $w_i = F(\alpha_i|\theta_1, \dots, \theta_k)$. The notation w_{Ti} represents the conjunctive probability of the background B with cause C_i. Similarly, w_{Tij} represents the conjunctive probability of the background B with causes C_i and C_j. By CARP's additivity assumption this probability is:

$$\begin{aligned} w_{Tij} &= F\left(\alpha_{Tij}|\theta_1, \dots, \theta_k\right) \\ &= F\left(\alpha_0 + \alpha_i C_1 + \alpha_j C_n|\theta_1, \dots, \theta_k\right) \end{aligned}$$

Note that this quantity also corresponds to the probability $w_{Tij} = P\left(e^+ \middle| c_i^+, c_j^+, b^+, \bigcap_{k \neq i,j} c_k^-\right)$. More generally, the probability of the effect when all causes are present is given by:

$$\begin{aligned} w_{T1\dots n} &= F(\alpha_{T1\dots n}|\theta_1, \dots, \theta_k) \\ &= F(\alpha_0 + \alpha_1 C_1 + \cdots + \alpha_n C_n|\theta_1, \dots, \theta_k) \end{aligned}$$

Binary C_i's can be represented with indicator functions. Some of the C_i's may also be continuous. In either case, the α_i represents the causal capacity contributed by a given cause C_i.

Now the challenge is that the response function θ_j parameters and the capacity α_i parameters are both unknown and estimates of each are conditional on the other. Markov chain Monte Carlo

(MCMC) methods can be applied to many problems of this type. The next section explores using the Metropolis-Hastings algorithm, an MCMC method. This section attempts the problem with a simple algorithm.

One option is to just maximize the likelihood over all the unknowns $\{\alpha_0, \alpha_1, \ldots, \alpha_n, \theta_1, \ldots, \theta_k\}$. Yet even for simple causal systems, this requires searching a large parameter space. To make the problem more tractable, an iterative algorithm similar to the coordinate descent approach is used. In coordinate descent, a complex multivariate problem is broken down into a sequence of lower dimensional optimization problems (Wright, 2015). The below approach is similar in that it breaks the problem into two optimization steps. The algorithm begins with starting coefficients $\boldsymbol{\alpha}^{(0)} = \left[\alpha_0^{(0)}, \alpha_1^{(0)}, \ldots, \alpha_n^{(0)}\right]$ and starting parameters $\boldsymbol{\theta}^{(0)} = \left\{\theta_0^{(0)}, \theta_1^{(0)}, \ldots, \theta_n^{(0)}\right\}$. On the first step, it conditions on the $\boldsymbol{\theta}^{(0)}$ density parameters and finds maximum likelihood estimates for the capacities, which are updated to $\boldsymbol{\alpha}^{(1)}$. Then on the second step, it conditions on $\boldsymbol{\alpha}^{(1)}$ and finds maximum likelihood density parameters $\boldsymbol{\theta}^{(1)}$. The first two steps are shown in the first two lines below:

$$\begin{aligned}
\boldsymbol{\alpha}^{(1)} &= \arg\max_{\boldsymbol{\alpha}} F\left(\boldsymbol{\alpha}\middle|\boldsymbol{y}, \boldsymbol{C}, \boldsymbol{\theta}^{(0)}\right) \\
\boldsymbol{\theta}^{(1)} &= \arg\max_{\boldsymbol{\theta}} F\left(\boldsymbol{\theta}\middle|\boldsymbol{y}, \boldsymbol{C}, \boldsymbol{\alpha}^{(1)}\right) \\
&\vdots \\
\boldsymbol{\alpha}^{(s)} &= \arg\max_{\boldsymbol{\alpha}} F\left(\boldsymbol{\alpha}\middle|\boldsymbol{y}, \boldsymbol{C}, \boldsymbol{\theta}^{(s-1)}\right) \\
\boldsymbol{\theta}^{(s)} &= \arg\max_{\boldsymbol{\theta}} F\left(\boldsymbol{\theta}\middle|\boldsymbol{y}, \boldsymbol{C}, \boldsymbol{\alpha}^{(s)}\right)
\end{aligned}$$

where $\boldsymbol{y}$ is a binary vector indicating whether the effect occurs and $\boldsymbol{C}$ is the design matrix with measurements for the causes. Note that the first column of $\boldsymbol{C}$ has all entries equal to 1 since it represents the influence of the always present context B. Also, the additive capacity assumption is reflected by the fact that $\boldsymbol{C}$ contains a single column for each cause with no interaction terms. The algorithm can be used iteratively to update the two sets of parameters until some convergence criterion is reached.

The response curves from Figure 5.10 were estimated using the simple algorithm from above. Specifically, the `R` code to simulate the learning data (for a given panel) is:

```
#population coefficients
a0 <- 0; a1 <- 0.2; a2 <- 0.5

#population shape parameters
sh1 <- 3; sh2 <- 3

Nobs <- 50
snum <- 3030
set.seed(snum)

#x1 will serve as the "context" variable while x2 is the candidate cause
x1 <- rnorm(Nobs) #scale variable
x1 <- rep(x1,2)
x2 <- rep(c(0,1),each=Nobs)

#transform x1 to [0,1] interval
tx1 <- (x1 - min(x1))/(max(x1) - min(x1))
```

And then the R code with the algorithm is given by:

```
#function to find max likelihood alpha coefficients
max.alphas <- function(aa, sh, ys){
  theta <- pbeta(aa[1] + aa[2]*tx1 + aa[3]*tx2, sh[1], sh[2])
  ptheta <- ifelse(ys==1, theta, 1-theta)
  LL <- -sum(log(ptheta))
  LL
}


#function to find max likelihood beta shape parameters
max.shapes <- function(sh, aa, ys){
  theta <- pbeta(aa[1] + aa[2]*tx1 + aa[3]*tx2, sh[1], sh[2])
  ptheta <- ifelse(ys==1, theta, 1-theta)
  LL <- -sum(log(ptheta))
  LL
}

#~~~~~~~~~~~#
# algorithm #
#~~~~~~~~~~~#

iterlim <- 100 #max number of iterations
crit <- c(.001, .1) #convergence criteria

#various starting values
E.shape_last <- c(1,1) #starting shape parameters
E.alpha_last <- rep(0.1,3) #starting coefs
E.shape <- E.shape_last #need this on first iteration

i <- 1

while(i <= iterlim){

  if(i > 1){
    E.alpha_last <- E.alpha
    E.shape_last <- E.shape
  }

  E.alpha <- nlminb(E.alpha_last, max.alphas, sh=E.shape, ys=ys)$par
```

```
  E.shape <- nlminb(E.shape_last, max.shapes, aa=E.alpha, ys=ys)$par

  par_diff <- max(abs(E.alpha - E.alpha_last))
  shape_diff <- max(abs(E.shape - E.shape_last))

  if(par_diff < crit[1] & shape_diff < crit[2]){
    print("convergence win")
    break
  }else i <- i+1; E.alpha_last <- E.alpha; E.shape_last <- E.shape; cat("*")

  if(i==iterlim){
    print("convergence fail")
  }

}
```

The simulation results shown in Chapter 5 suggest that this algorithm does a decent job. It is also relatively fast, typically producing estimates in a matter of seconds. A disadvantage is that its properties regarding convergence are unknown. Superficial inspection also suggests that the algorithm is a high variance estimator of response curves.

I.2 Metropolis algorithm

The estimation problem may also be cast within a Bayesian framework. For many Bayesian models it is difficult or impossible to derive conjugate or semi-conjugate distributions. The Metropolis algorithm is a general method that can be brought to bear on such problems. In the Metropolis algorithm, parameters are randomly sampled from a "proposal distribution". The likelihood and prior distributions are then used to determine whether the proposal is accepted. The resultant parameter sequence forms a Markov chain that will eventually converge to the posterior distribution (see Hoff (2009) Chapter 10 for an explanation of the Metropolis-Hastings algorithm). Thus, samples from the Markov chain can be used to approximate the posterior distribution. An advantage of the Bayesian approach is that uncertainty estimates for the parameters are easily obtained as posterior confidence intervals.

To see how this can work, suppose a dichotomous effect y that is caused by either the background cause B, a continuous cause C_1, or a dichotomous cause C_2. Further, suppose a two-parameter response function maps capacity to outcome probability:

$$\Pr(y_i = 1) = F\left(\alpha_0 + \alpha_1 C_{1(i)} + \alpha_2 C_{2(i)} \middle| \alpha_0, \alpha_1, \alpha_2, \theta_1, \theta_2\right) \qquad \text{(I.1)}$$

where α_0, α_1 and α_2 are the capacities while the response function is determined by parameters θ_1 and θ_2. We are interested in finding the joint posterior distribution $p(\alpha_0, \alpha_1, \alpha_2, \theta_1, \theta_2 | \boldsymbol{y}, \boldsymbol{C})$. We can assume a prior distribution for the parameters, so we only need to form the likelihood, which is easily obtained. For a given set of parameters, equation (I.1) implies a probability $\Pr(y_i = 1 | \alpha_0, \alpha_1, \alpha_2, \theta_1, \theta_2) = \omega_i$. The likelihood is just then:

$$\mathcal{L}(\alpha_0, \alpha_1, \alpha_2, \theta_1, \theta_2 | \boldsymbol{y}, \boldsymbol{C}) = \prod_{i=1}^{n} \omega_i^{y_i} (1 - \omega_i)^{1 - y_i}$$

The prior and likelihood can then be used within the Metropolis algorithm to approximate the posterior distribution.

I.2.1 Simulation example

For a concrete example, assume that C_1 is continuous and normally distributed while C_2 is dichotomous. Measurements of cause C_1 are mapped to the $[0,1]$ interval using $\mathrm{c}_1' = [\mathrm{c}_1 - \min(\mathrm{c}_1)]/[\max(c_1) - \min(c_1)]$. This mapping ensures that the α coefficients for C_1 and C_2 will be similar in magnitude. Further, assume the true parameter values are $\alpha_0 = 0$, $\alpha_1 = 0.2$, $\alpha_2 = 0.5$ and $\theta_1 = 3$ and $\theta_2 = 1$ (note that the α_1 parameter is for the transformed c_1').

Two simulations are performed using the above population parameters. For the first simulation 20 observations are randomly sampled for the continuous C_1. At each of these 20 observations, 10 observations of the dichotomous C_2 are randomly sampled, evenly split between cause present and cause absent. This makes for a total of $N = 200$ observations. The second simulation uses a larger sample, with 200 observations of the C_1 cause and 20 observations of C_2 at each value of C_1, producing a total of $N = 4000$ observations.

A uniform prior is assumed for all the α coefficient parameters while a $\mathrm{Gamma}(1,1)$ prior is used for each of the θ_1, θ_2 shape parameters. All parameters are assumed independent in the prior. A uniform proposal distribution is used for both sets of parameters (the `R` code appended below shows specifically how this was done). The Metropolis algorithm was run for 100,000 iterations, and the first 5000 were discarded as the "burn-in" sample. Plots for the two simulations are shown in Figure I.1. Both of the estimated response curves capture the general qualitative relationship. Both curves also appear to give decent estimates for low to moderate capacity. As capacity

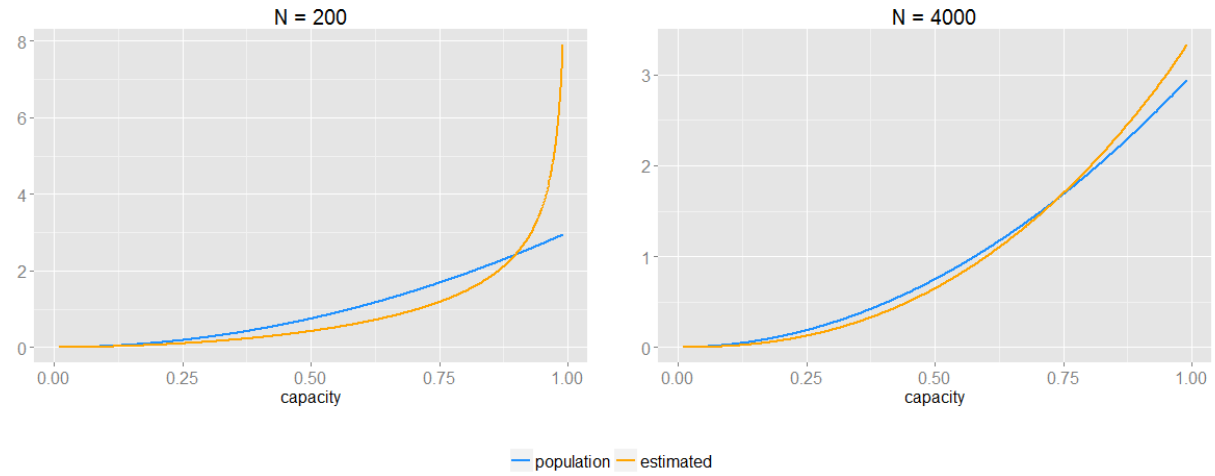


Figure I.1. Population (blue) and estimated (orange) response curves for training samples of size N=200 (left panel) and size N=4000 (right panel). Population response curve is beta(3,1) density. The posterior distribution is approximated with a Metropolis algorithm of 10,000 iterations, with first 5000 discarded as the burn-in sample. Posterior expectations are used for the estimated shape parameters.

increases there is more curvature in the true relationship, making estimation more difficult. Summary statistics from the two simulations are presented in Table I.1. Estimates from the larger sample size are generally more precise (with the exception of the θ_2 shape parameter). Yet even for $N = 4000$, there is considerable uncertainty in the estimates.

Table I.1. Posterior expectations and 95% confidence intervals estimated from Metropolis algorithm samples.

	N=200		*N=4000*	
Parameter	Posterior Expectation	Posterior 95% confidence interval	Posterior Expectation	Posterior 95% confidence interval
α_0	0.09	[0.00, 0.31]	0.06	[0.00, 0.20]
α_1	0.22	[0.03, 0.42]	0.19	[0.11, 0.27]
α_2	0.57	[0.34, 0.81]	0.51	[0.31, 0.71]
θ_1	2.80	[1.34, 5.21]	3.36	[2.13, 5.02]
θ_2	0.57	[0.15, 1.48]	0.99	[0.19, 2.63]

Notes. Results of *N*=200 shown in left panel and *N*=4000 shown in the right panel. For both *N*'s, the Metropolis algorithm was iterated 100,000 times. The first 5000 samples were omitted from the chain as the burn-in sample.

The Metropolis algorithm constructed for this problem is quite basic. Improvements will need to be made in order to obtain good estimates with a reasonable amount of data. The R code for the algorithm is shown below (simulation data for the function may be generated from the R code shown in Section I.1):

```
#Function to estimate response curves from two causes (predictors) by using
the Metropolis algorithm.

#arguments

# y     data vector of 0/1 outcomes
# x1    data vector of measurements for first cause
# x2    data vector of measurements for second cause
# sim   number of iterations for the chain

# alpha_0s     starting values for the three alpha coefficients
# shape_0s     starting values for the two beta shape parameters

# delta1     half the width of the proposal distribution for the alpha vector
# delta2     half the width of the proposal distribution for the shape vector

# prior.bs     specifies priors for the coefficients. Default is beta(1,1)
# prior.sh     specifies priors for the beta shapes. Default is gamma(1,1)

emp_mh <- function(y, x1, x2, sim=100000,
                   alpha_0s=rep(0.1,3), shape_0s=c(1,1),
                   delta1 = .05, delta2 = .1,
                   prior.as=c(1,1), prior.sh=c(1,1) ){

  aa <- alpha_0s; sh <- shape_0s
  CHAIN <- matrix(data=NA, nrow=sim, ncol=5)

  for(i in 1:sim){

    aa.star <- aa + runif(3, -delta1, delta1)
    sh.star <- sh + runif(2, -delta2, delta2)

    #enforce proper values for the shape parameters
    sh.star <- ifelse(sh.star <= 0, c(.01, .01), sh.star)

    theta <- pbeta(aa[1] + aa[2]*x1 + aa[3]*x2, sh[1], sh[2])
    theta.star <- pbeta(aa.star[1] + aa.star[2]*x1 + aa.star[3]*x2,
                  sh.star[1], sh.star[2])

    lp <- sum(dbinom(y, 1, theta, log=TRUE)) +
      sum(dbeta(aa, prior.as[1], prior.as[2], log=TRUE)) +
      sum(dgamma(sh, prior.sh[1], prior.sh[2], log=TRUE))

    lp.star <- sum(dbinom(y, 1, theta.star, log=TRUE)) +
      sum(dbeta(aa.star, prior.as[1], prior.as[2], log=TRUE)) +
      sum(dgamma(sh.star, prior.sh[1], prior.sh[2], log=TRUE))

    log.r <- lp.star - lp

    if(log(runif(1)) < log.r){ aa <- aa.star; sh <- sh.star}
    CHAIN[i,] <- c(aa, sh)
    }
```

```
  CHAIN <- data.frame(1:sim, CHAIN)
  names(CHAIN) <- c("sim", "a0", "a1", "a2", "sh1", "sh2")
  CHAIN
}
```

References

Abbott, J. T., Hamrick, J. B., & Griffiths, T. L. (2013). Approximating Bayesian inference with a sparse distributed memory system. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 1686–1691.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, *114*(3), 435–448. https://doi.org/10.1037/0033-2909.114.3.435

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*(4), 381–405. https://doi.org/10.1016/0023-9690(83)90024-3

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Anderson, J. R. (1991a). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–517. https://doi.org/10.1017/S0140525X00070801

Anderson, J. R. (1991b). The Adaptive Nature of Human Categorization. *Psychological Review*, *98*(3), 409–29.

Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition*, *23*(4), 510–524. https://doi.org/10.3758/BF03197251

Arkes, H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General*, *112*(1), 117–135. https://doi.org/10.1037/0096-3445.112.1.117

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as Probability Density Estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233. https://doi.org/10.1006/jmps.1995.1021

Baker, A. G., Vallée-Tourangeau, F., & Murphy, R. A. (2000). Asymptotic judgment of cause in a relative validity paradigm. *Memory & Cognition Memory & Cognition*, *28*(3), 466–479.

Barkow, J. H., Cosmides, L., & Tooby, J. (1992). *The Adapted mind: evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Bes, B., Sloman, S., Lucas, C. G., & Raufaste, É. (2012). Non-Bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*(7), 1178–1203. https://doi.org/10.1111/j.1551-6709.2012.01262.x

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*(1), 220–251. https://doi.org/10.1037/a0014462

Bowers, J. S., & Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414. https://doi.org/10.1037/a0026450

Bowers, J. S., & Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin*, *138*(3), 423–426. https://doi.org/10.1037/a0027750

Brighton, H., & Gigerenzer, G. (2008). Bayesian brains and cognitive mechanisms: harmony or dissonance? *The Probabilistic Mind : Prospects for Bayesian Cognitive Science, 189-208 (2008)*.

Buehner, M., & Cheng, P. (1997). {Causal induction: The power PC theory versus the Rescorla-Wagner model} (p. 55). Presented at the Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society: August 7-10, 1997, Stanford University, Lawrence Erlbaum Associates.

Buehner, M., Cheng, P. W., & Clifford, D. (2003). From Covariation to Causation: A Test of the Assumption of Causal Power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1119–1140. https://doi.org/10.1037/0278-7393.29.6.1119

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory, Vol. 1: Cognition; Vol. 2: Social; Vol. 3: Developmental.* (pp. 187–215). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford; New York: Clarendon Press ; Oxford University Press.

Cartwright, N. (2007). *Causal Powers: What are They? Why Do We Need Them? What Can and Cannot be Done with Them?* Contingency and Dissent in Science Project.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *The New England Journal of Medicine*, *299*(18), 999–1001.

Cavagnaro, D. R., Pitt, M. A., Gonzalez, R., & Myung, J. I. (2013). Discriminating among probability weighting functions using adaptive design optimization. *Journal of Risk and Uncertainty*, *47*(3), 255–289. https://doi.org/10.1007/s11166-013-9179-3

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, *18*(5), 537–545.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. https://doi.org/10.1037/0033-295X.104.2.367

Cheng, P. W. (2000). Causality in the mind: Estimating contextual and conjunctive power. In F. C. Keil, R. A. Wilson, F. C. (Ed) Keil, & R. A. (Ed) Wilson (Eds.), *Explanation and cognition.* (pp. 227–253). Cambridge, MA, US: The MIT Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. https://doi.org/10.1177/001316446002000104

Collins, D. J., & Shanks, D. R. (2006). Conformity to the power PC theory of causal induction depends on the type of probe question. *The Quarterly Journal of Experimental Psychology*, *59*(2), 225–232. https://doi.org/10.1080/17470210500370457

Croen, L. A., Najjar, D. V., Fireman, B., & Grether, J. K. (2007). Maternal and paternal age and risk of autism spectrum disorders. *Archives of Pediatrics & Adolescent Medicine*, *161*(4), 334–340. https://doi.org/10.1001/archpedi.161.4.334

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121. https://doi.org/10.1016/S0022-2496(02)00016-0

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, 59–75.

Danks, D. (2014). The Mathematics of Causal Capacities. Retrieved from http://philsci-archive.pitt.edu/11113/

Danks, D., & Eberhardt, F. (2011). Keeping Bayesian models rational: The need for an account of algorithmic rationality. *Behavioral and Brain Sciences*, *34*(04), 197–197. https://doi.org/10.1017/S0140525X11000240

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical Causal Learning. *Advances in Neural Information Processing Systems.*, (15), 83–90.

Dieckmann, A., & Rieskamp, J. (2007). The influence of information redundancy on probabilistic inferences. *Memory & Cognition*, *35*(7), 1801–1813. https://doi.org/10.3758/BF03193511

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines*, *21*(3), 389–410. https://doi.org/10.1007/s11023-011-9241-3

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3–19. https://doi.org/10.1037/0033-2909.99.1.3

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159–176. https://doi.org/10.1016/j.cognition.2012.11.014

Endress, A. D. (2014). How are Bayesian models really used? Reply to Frank (2013). *Cognition*, *130*(1), 81–84. https://doi.org/10.1016/j.cognition.2013.09.003

Fales, E., & Wasserman, E. A. (1992). Causal knowledge: What can psychology teach philosophers? *Journal of Mind and Behavior*, *13*(1), 1–27.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 678–693. https://doi.org/10.1037/a0014928

Frank, M. C. (2013). Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013). *Cognition*, *128*(3), 417–423. https://doi.org/10.1016/j.cognition.2013.04.010

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371. https://doi.org/10.1016/j.cognition.2010.10.005

Geisler, W. S. (1989). Ideal observer theory in psychophysics and physiology. *Physica Scripta*, *39*(1), 153–160. https://doi.org/10.1088/0031-8949/39/1/025

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771–781. https://doi.org/10.1016/j.visres.2010.09.027

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1–24. https://doi.org/10.1162/NECO_a_00226

Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases." *European Review of Social Psychology*, *2*(1), 83–115. https://doi.org/10.1080/14792779143000033

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: the adaptive toolbox*. Cambridge, Mass.: MIT Press.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart.* (pp. 3–34). New York, NY, US: Oxford University Press.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, *8*(1), 39–60. https://doi.org/10.1023/A:1008234330618

Glymour, C. (2000). Bayes nets as psychological models. In F. C. Keil, R. A. Wilson, F. C. (Ed) Keil, & R. A. (Ed) Wilson (Eds.), *Explanation and cognition.* (pp. 169–197). Cambridge, MA, US: The MIT Press.

Glymour, C. (2002). The Mind's Arrows. Bayes Nets and Graphical Causal Models in Psychology: Book review. *Acta Psychologica*, *111*(3), 355–357. https://doi.org/10.1016/S0001-6918(02)00058-6

Goodman, N. D., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P. W., & Hamrick, J. B. (2015). Relevant and robust: A response to Marcus and Davis (2013). *Psychological Science*, *26*(4), 539–541. https://doi.org/10.1177/0956797614559544

Green, D. M. (1960). Psychoacoustics and Detection Theory. *The Journal of the Acoustical Society of America*, *32*, 1189. https://doi.org/10.1121/1.1907882

Green, D. M. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*(3), 415–422. https://doi.org/10.1037/a0026884

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive bases. *Cognitive Science*, *32*(1), 68–107. https://doi.org/10.1080/03640210701801974

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480. https://doi.org/10.1080/15326900701326576

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun & R. (Ed) Sun (Eds.), *The Cambridge handbook of computational psychology*. (pp. 59–100). New York, NY, US: Cambridge University Press.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. https://doi.org/10.1111/tops.12142

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. https://doi.org/10.1016/j.cogpsych.2005.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal Predictions in Everyday Cognition. *Psychological Science*, *17*(9), 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. https://doi.org/10.1177/0963721412447619

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal Reasoning Through Intervention. In A. Gopnik & L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press.

Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psychology*, *5*.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, *39*(12), 1372–1388. https://doi.org/10.1037/0003-066X.39.12.1372

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* (2nd ed.). Dordrecht: Springer.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York; London: Springer.

Holyoak, K. J., & Cheng, P. W. (2011). Causal Learning and Inference as a Rational Process: The New Synthesis. *Annual Review of Psychology*, *62*, 135.

Holyoak, K. J., & Lu, H. (2011). What the Bayesian framework has contributed to understanding cognition: Causal learning as a case study. *Behavioral and Brain Sciences*, *34*(4), 203–204. https://doi.org/10.1017/S0140525X1100032X

Hume, D. (1854). *The philosophical works of David Hume*. Boston: Little, Brown and company ; Edinburgh: Little, Brown and company.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: an essay on the construction of formal operational structures*. [New York]: Basic Books.

James, G. M., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An introduction to statistical learning: with applications in R* ([Corrected at 6th printing 2015].). New York: Springer : Springer Science+Business Media.

Jaynes, E. T. (2003). *Probability Theory The Logic of Science*. Cambridge: Cambridge University Press.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17. https://doi.org/10.1037/h0093874

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(04), 169–188. https://doi.org/10.1017/S0140525X10003134

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed.). New York: Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1982). Evidential impact of base rates. *Judgement under Uncertainty: Heuristics and Biases*, 153–160.

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1363–1386. https://doi.org/10.1037/0278-7393.19.6.1363

Katsikopoulos, K. V., & Martignon, L. (2006). Naïve heuristics for paired comparisons: Some results on their relative accuracy. *Journal of Mathematical Psychology*, *50*(5), 488–494. https://doi.org/10.1016/j.jmp.2006.06.001

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. https://doi.org/10.1037/0033-295X.94.2.211

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*. New York Academic Press.

Kwisthout, J., Wareham, T., & Van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, *35*(5), 779–84. https://doi.org/10.1111/j.1551-6709.2011.01182.x

Lee, M. D., & Zhang, S. (2012). Evaluating the coherence of Take-the-best in structured environments. *Judgment and Decision Making*, *7*(4), 360.

Lehmann, E. L., Casella, G., & Fienberg, S. (1998). *Theory of Point Estimation*. Secaucus: Springer-Verlag New York, Incorporated.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*(6), 493–497. https://doi.org/10.1016/S0960-9822(03)00135-0

Levin, I. P., Wasserman, E. A., & Kao, S. (1993). Multiple methods for examining biased information use in contingency judgments. *Organizational Behavior and Human Decision Processes*, *55*(2), 228–250. https://doi.org/10.1006/obhd.1993.1032

Levy, R. P., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 937–944). Curran Associates, Inc.

Liljeholm, M., & Cheng, P. W. (2009). The influence of virtual sample size on confidence and causal-strength judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 157–172. https://doi.org/10.1037/a0013972

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, *107*(1), 195.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage Publications.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2006). Modeling Causal Learning Using Bayesian Generic Priors on Generative and Preventive Powers. *Department of Statistics, UCLA*. Retrieved from http://escholarship.org/uc/item/29r5b829

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984. https://doi.org/10.1037/a0013256

Lu, H., Yuille, A. L., Liljiholm, M., Cheng, P. W., & Holyoak, K. J. (2007). Bayesian Models of Judgments of Causal Strength: A Comparison. *eScholarship*. Retrieved from http://escholarship.org/uc/item/7mx3b8rg

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432. https://doi.org/10.1038/nn1790

Marcus, G. F. (2008). *Kluge: the haphazard construction of the human mind*. Boston: Houghton Mifflin.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science, 24*(12), 2351–2360. https://doi.org/10.1177/0956797613495418

Markman, A., & Otto, A. (2011). Cognitive systems optimize energy rather than information. *Behavioral and Brain Sciences*, *34*(4), 207–207. https://doi.org/10.1017/S0140525X11000355

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In *Simple heuristics that make us smart.* (pp. 119–140). New York, NY, US: Oxford University Press.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*(1), 29–71. https://doi.org/10.1023/A:1015516217425

Mcclelland, J. L., Mcnaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33–61. https://doi.org/10.1016/j.cogpsych.2006.04.004

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*(7), 1133–1147. https://doi.org/10.1080/03640210802353016

Newell, A. (1982). *The knowledge level*. Pittsburgh, Pa.: Design Research Center, Carnegie-Mellon University.

Nisbett, R. E., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgment*. Prentice-Hall.

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Exemplar-Based Accounts of Relations Between Classification, Recognition, and Typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Novick, L. R., & Cheng, P. W. (2004). Assessing Interactive Causal Influence. *Psychological Review*, *111*(2), 455–485. https://doi.org/10.1037/0033-295X.111.2.455

Oaksford, M., & Chater, N. (2007). *Bayesian rationality the probabilistic approach to human reasoning*. New York, NY, US: Oxford University Press.

Oaksford, M., & Chater, N. (2010). *Cognition and conditionals: probability and logic in human thinking*. Oxford ; New York: Oxford University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*(1), 61–73. https://doi.org/10.1037/0033-295X.94.1.61

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2nd ed.). Cambridge ; New York: Cambridge University Press.

Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology Section A*, *56*(6), 977–1007. https://doi.org/10.1080/02724980244000738

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, *14*(4), 577–596. https://doi.org/10.3758/BF03196807

Perales, J. C., & Shanks, D. R. (2008). Driven by power? Probe question and presentation format effects on causal judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1482–1494. https://doi.org/10.1037/a0013509

Pollitt, E., Cueto, S., & Jacoby, E. R. (1998). Fasting and cognition in well- and undernourished schoolchildren: a review of three experimental studies.(Breakfast, Cognition, and School Learning: Proceedings of a Symposium Held in Napa,California, August 28-30, 1995.). *American Journal of Clinical Nutrition*, *67*(4), 779S.

Pylyshyn, Z. W. (1984). *Computation and cognition: toward a foundation for cognitive science*. Cambridge, Mass.: MIT Press.

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *ResearchGate, Vol. 2*.

Russell, P. J. (2003). *Essential iGenetics*. San Francisco: Benjamin Cummings.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. https://doi.org/10.1037/a0020511

Sandin, S., Schendel, D., Magnusson, P., Hultman, C., Surén, P., Susser, E., … Reichenberg, A. (2015). Autism risk associated with parental age and with increasing difference in age between the parents. *Molecular Psychiatry*. https://doi.org/10.1038/mp.2015.70

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, *43*(5), 1124–1139. https://doi.org/10.1037/0012-1649.43.5.1124

Sen-Chowdhry, S., & McKenna, W. J. (2006). Sudden Cardiac Death in the Young: A Strategy for Prevention by Targeted Evaluation. *Cardiology*, *105*(4), 196–206. https://doi.org/10.1159/000091640

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory & Cognition*, *8*(5), 459–467. https://doi.org/10.3758/BF03211142

Shanks, D. R. (1985). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition*, *13*(2), 158–167. https://doi.org/10.3758/BF03197008

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation*, *18*(2), 147–166. https://doi.org/10.1016/0023-9690(87)90008-7

Shanks, D. R. (1995). Is human learning rational? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *48A*(2), 257–279. https://doi.org/10.1080/14640749508401390

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower & G. H. (Ed) Bower (Eds.), *The psychology of learning and motivation: Advances in research and theory, Vol. 21.* (pp. 229–261). San Diego, CA, US: Academic Press.

Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing Associative and Probabilistic Contrast Theories of Human Contingency Judgment. *Psychology of Learning and Motivation*, *34*, 265–311. https://doi.org/10.1016/S0079-7421(08)60563-0

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464. https://doi.org/10.3758/PBR.17.4.443

Shou, Y., & Smithson, M. (2015). Effects of question formats on causal judgments and model evaluation. *Frontiers in Psychology*, *6*, 467. https://doi.org/10.3389/fpsyg.2015.00467

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London ; New York: Chapman and Hall.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Sloman, S., & Fernbach, P. M. (2008). The value of rational analysis: an assessment of causal reasoning and learning.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology*, *4*(3), 165–173. https://doi.org/10.1111/j.1467-9450.1963.tb01324.x

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

Swets, J., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–40.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure Learning in Human Causal Induction. In *IN* (pp. 59–65). MIT Press.

Todd, P. M., & Gigerenzer, G. (2007). Environments That Make Us Smart: Ecological Rationality. *Current Directions in Psychological Science*, *16*(3), 167.

Vallée-Tourangeau, F., Murphy, R. A., Drew, S., & Baker, A. G. (1998). Judging the importance of constant and variable candidate causes: A test of the power PC theory. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *51A*(1), 65–84. https://doi.org/10.1080/027249898391765

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, *25*(2), 127–151. https://doi.org/10.1006/lmot.1994.1008

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Wagner, A. R. (1969). Stimulus Selection and A "Modified Continuity Theory." In G. H. B. and J. T. Spence (Ed.), *Psychology of Learning and Motivation* (Vol. 3, pp. 1–41). Academic Press.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *19*(3), 231–241. https://doi.org/10.1037/h0082908

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response–outcome contingencies under free-operant procedures. *Learning and Motivation*, *14*(4), 406–432. https://doi.org/10.1016/0023-9690(83)90025-5

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 509–521. https://doi.org/10.1037/0278-7393.16.3.509

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 174–188. https://doi.org/10.1037/0278-7393.19.1.174

White, P. A. (2003). Causal judgment as evaluation of evidence: the use of confirmation and disconfirmatory information. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *56*(3), 491–513.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: substantially improving power and accuracy*. New York ; London: Springer. Retrieved from http://dx.doi.org/10.1007/978-1-4419-5525-8

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Amsterdam ; Boston: Academic Press.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York: Springer.

Wright, S. (2015). Coordinate descent algorithms. *Mathematical Programming*, *151*(1), 3–34. https://doi.org/10.1007/s10107-015-0892-3

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, *76*, 1–29. https://doi.org/10.1016/j.cogpsych.2014.11.001