ELSEVIER

Structure and strength in causal induction ☆

Thomas L. Griffiths [a],*, Joshua B. Tenenbaum [b]

[a] *Department of Psychology, Stanford University, USA*
[b] *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology,
77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA*

Abstract

We present a framework for the rational analysis of elemental causal induction—learning about the existence of a relationship between a single cause and effect—based upon causal graphical models. This framework makes precise the distinction between causal structure and causal strength: the difference between asking whether a causal relationship exists and asking how strong that causal relationship might be. We show that two leading rational models of elemental causal induction, ΔP and causal power, both estimate causal strength, and we introduce a new rational model, causal support, that assesses causal structure. Causal support predicts several key phenomena of causal induction that cannot be accounted for by other rational models, which we explore through a series of experiments. These phenomena include the complex interaction between ΔP and the base-rate probability of the effect in the absence of the cause, sample size effects, inferences from incomplete contingency tables, and causal

learning from rates. Causal support also provides a better account of a number of existing datasets than either ΔP or causal power.

> The contagion spread rapidly and before its progress could be arrested, sixteen persons were affected of which two died. Of these sixteen, eight were under my care. On this occasion I used for the first time the affusion of cold water, in the manner described by Dr. Wright. It was first tried in two cases... The effects corresponded exactly with those mentioned by him to have occurred in his own case and thus encouraged the remedy was employed in five other cases. It was repeated daily, and of these seven patients, the whole recovered. (James Currie, 1798/1960, p. 430)

1. Introduction

Statistical methods for evaluating the relationships between variables were only developed at the end of the 19th century, more than 200 years after the birth of modern science. The foundations of physics, chemistry, biology, and medicine were all laid before formal methods for analyzing correlations or contingency tables existed. Even difficult statistical problems like evaluating medical treatments were addressed by early scientists, as the epigraph illustrates. Its author, Dr. James Currie, was an 18th century ship's surgeon who later went into practice in Liverpool. After having heard a Dr. William Wright give an account of the efficacy of being doused with cold water in treating an extended fever, Currie conducted his own experiment, with results described above. He was sufficiently encouraged that he went on to use the treatment with hundreds of other patients, publishing a detailed treatise on the matter (Currie, 1798/1960). Washing the skin of the patient is still used to ease fevers, although modern medicine cautions against using water cold enough to induce shivering.

While the development of statistical methods for designing and analyzing experiments has greatly streamlined scientific argument, science was possible before statistics: in many cases, the causal relationships between variables that are critical to understanding our world could be discovered without any need to perform explicit calculations. Science was possible because people have a capacity for causal induction, inferring causal structure from data. This capacity is sufficiently accurate as to have resulted in genuine scientific discoveries, and provides the basis for the construction of the intuitive theories that express our knowledge about the world. Currie's assessment of the water treatment is directly analogous to the kinds of causal inferences we perform every day, such as evaluating whether taking vitamins prevents us from getting sick, or whether drinking coffee increases our productivity.

The most basic problem of causal induction is learning that a relationship exists between a single cause and effect. This problem of *elemental causal induction* has been

the subject of most previous studies of human causal judgment. This simplest case is sufficiently constrained to allow rigorous testing of mathematical models against people's judgments of cause–effect relations. Recent accounts of elemental causal induction have emphasized the rational basis of human learning, presenting formal analyses of how an agent should learn about causal relationships (e.g., Anderson, 1990; Cheng, 1997; López, Cobos, Caño, & Shanks, 1998). This strategy has resulted in several distinct mathematical models of causal judgment, none of which explains all of the phenomena of causal induction. Consequently, there is an ongoing debate about which model gives a better account of human judgments (e.g., Cheng, 1997; Lober & Shanks, 2000).

In this paper, we present a framework for analyzing the computational problem of elemental causal induction, based on causal graphical models. Causal graphical models are a set of tools for learning and reasoning about causal relationships that has been developed by computer scientists, statisticians, and philosophers (e.g. Pearl, 2000; Spirtes, Glymour, & Schienes, 1993). Our framework clarifies the assumptions made by previous rational models, such as ΔP (Allan, 1980; Jenkins & Ward, 1965; López et al., 1998) and causal power (Cheng, 1997), and results in a fundamentally different model of human judgments, which we call "causal support." While previous models view human judgments as reflecting the *strength* of a causal relationship, causal support addresses the *structural* question of whether or not a causal relationship exists.

Causal graphical models have recently become the focus of a great deal of interest among cognitive scientists, with several studies examining the extent to which human causal reasoning corresponds to the qualitative assumptions of this approach (Danks & McKenzie, submitted; Gopnik et al., 2004; Glymour, 1998, 2001; Lagnado & Sloman, 2002; Waldmann & Martignon, 1998). Our work extends these results by showing that causal graphical models can be used to make quantitative predictions about human behavior (e.g., Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Tenenbaum & Griffiths, 2001, 2003). The framework we present here uses causal graphical models to explicate the roles of structure and strength in the problem of causal induction, and in defining causal support, our model of human judgments.

Causal support predicts several phenomena that are problematic for other rational models. Our presentation will be organized around these phenomena. The first phenomenon we will consider is the interaction between covariation, measured by ΔP, and the base-rate probability of the effect in the absence of the cause in determining human judgments. This interaction manifests in two curious phenomena, the "frequency illusion"—a decrease in causal judgments as the base-rate decreases when $\Delta P = 0$ (Allan & Jenkins, 1983; Buehner, Cheng, & Clifford, 2003; Shanks, López, Darby, & Dickinson, 1996)—and non-monotonic effects of changes in base-rate at other values of ΔP (Lober & Shanks, 2000). We will also discuss effects of sample size (White, 1998, 2002c, 2003c), inferences from incomplete contingency tables, and causal induction from rates (c.f. Anderson & Sheu, 1995; Wasserman, 1990). No other rational model of can explain all of these phenomena, or fit as wide a range of datasets as causal support.

The plan of the paper is as follows. First we outline the problem of elemental causal induction in more detail, describing the experimental paradigms that are

Table 1
Contingency table representation used in elemental causal induction

	Effect present (e^+)	Effect absent (e^-)
Cause present (c^+)	$N(e^+, c^+)$	$N(e^-, c^+)$
Cause absent (c^-)	$N(e^+, c^-)$	$N(e^-, c^-)$

the focus of our investigation, the two leading rational models, ΔP and causal power, and some of the data that has been gathered in support of them. Then, we provide a brief summary of causal graphical models, present our framework for analyzing the problem of elemental causal induction, and use this to derive causal support. The body of the paper discusses the phenomena predicted by causal support but not by other models, explaining the statistical origins of these predictions. We close by considering the circumstances under which we expect causal support to be most consistent with human judgments, its relationship with ideas such as ''reliability'' (Buehner & Cheng, 1997; Perales & Shanks, 2003), and how this account of elemental causal induction can be extended to shed light on other aspects of causal learning.

2. Elemental causal induction

Much psychological research on causal induction has focused upon the problem of learning a single causal relation: given a candidate cause, C, and a candidate effect, E, people are asked to assess the relationship between C and E.[1] Most studies present information corresponding to the entries in a 2×2 contingency table, as in Table 1. People are given information about the frequency with which the effect occurs in the presence and absence of the cause, represented by the numbers $N(e^+, c^+), N(e^-, c^-)$ and so forth. In a standard example, C might be injecting a chemical into a mouse, and E the expression of a particular gene. For this case, $N(e^+, c^+)$ would be the number of injected mice expressing the gene, while $N(e^-, c^-)$ would be the number of uninjected mice not expressing the gene.

This contingency information is usually presented to participants in one of three modes. Early experiments on causal induction would either explicitly provide participants with the numbers contained in the contingency table (e.g., Jenkins & Ward, 1965), which we will refer to as a ''summary'' format, or present individual cases one by one, with the appropriate frequencies (e.g., Ward & Jenkins, 1965), which we will refer to as an ''online'' format. Some more recent experiments use a mode of presentation between these two extremes, showing a list of all individual cases simultaneously (e.g., Buehner et al., 2003; White, 2003c), which we will refer to as a ''list'' format.

---

[1] We will represent variables such as C, E with capital letters, and their instantiations with lowercase letters, with c^+, e^+ indicating that the cause or effect is present, and c^-, e^- indicating that the cause or effect is absent.

Experiments also differ in the questions that are asked of the participants. Participants can be asked to rate the strength of the causal relationship, the probability of a causal relationship, or their confidence that a causal relationship exists. Understanding the effects of question wording is an ongoing task (e.g., White, 2003b), but one variable that has been shown to have a strong effect is asking counterfactual questions, such as "What is the probability that a mouse not expressing the gene before being injected will express it after being injected with the chemical?" (Buehner et al., 2003; Collins & Shanks, submitted).

Causal induction tasks also vary in their treatment of the valence of the potential cause, and the nature of the rating scale used for responses. Causes can be either "generative," increasing the probability of an outcome (as in our mouse gene example), or "preventive," reducing its probability (as in the case of Dr. Currie's cold water treatment). Some experiments use exclusively generative or exclusively preventive causes and ask for judgments on a nonnegative scale (e.g., 0–100), while others mix generative and preventive causes and ask for judgments on a scale that has both positive and negative ends (e.g., −100 to 100).

Given the many ways in which experiments on causal judgment can differ, it is important to identify the scope of the present analysis. We will discuss experiments that use all three modes of presentation, as each mode captures an aspect of causal induction that is important for the development of rational models: the summary format removes memory demands and allows a deliberative inference, the online format taps intuitions about causality that are engaged by direct interaction with data, and the list format falls between these extremes. We will focus on experiments that require participants to make judgments about potential causes of a single kind, generative or predictive. Most of the critical datasets in the current debate about rational models of causal induction are of this form (e.g., Buehner & Cheng, 1997; Lober & Shanks, 2000). In Section 15, we will consider how our framework can be extended to shed light on other issues in causal induction, including learning about multiple potential causes, the dynamics of causal judgments in online tasks, and combining generative and preventive causes.

2.1. Rational models of elemental causal induction

In the spirit of Marr's (1982) computational level and Anderson's (1990) rational analysis, recent theories have tried to establish the task of causal induction as a computational problem and use the optimal solution to that problem to explain human behavior. We will describe two leading rational models of causal induction which are at the center of a debate about modeling causal judgments: ΔP and causal power.

2.1.1. ΔP and associative strength

One common approach to modeling judgments about causal relationships is to combine the frequencies from a contingency table in the form

$$\Delta P = \frac{N(e^+, c^+)}{N(e^+, c^+) + N(e^-, c^+)} - \frac{N(e^+, c^-)}{N(e^+, c^-) + N(e^-, c^-)} = P(e^+|c^+) - P(e^+|c^-),$$

(1)

where $P(e^+|c^+)$ is the empirical conditional probability of the effect given the presence of the cause, estimated from the contingency table counts $N(\cdot)$. ΔP thus reflects the change in the probability of the effect occuring as a consequence of the occurence of the cause. This measure was first suggested by Jenkins and Ward (1965), subsequently explored by Allan (Allan, 1980, 1993; Allan & Jenkins, 1983), and has appeared in various forms in both psychology and philosophy (Cheng & Holyoak, 1995; Cheng & Novick, 1990, 1992; Melz, Cheng, Holyoak, & Waldmann, 1993; Salmon, 1980). One argument for the appropriateness of ΔP as a normative model uses the fact that it is the asymptotic value of the weight given to the cause C when the causal induction task is modeled with a linear associator trained using the Rescorla–Wagner (Rescorla & Wagner, 1972) learning rule (Cheng, 1997; Cheng & Holyoak, 1995; Chapman & Robbins, 1990; Danks, 2003; Wasserman, Elek, Chatlosh, & Baker, 1993).

2.1.2. The power PC theory and causal power

Cheng (1997) rejected ΔP as a measure of causal strength because it is a measure of covariation, not causality. According to Cheng (1997) and Novick and Cheng (2004), human judgments reflect a set of assumptions about causality that differ from those of purely "covariational" measures such as ΔP and conventional statistics. Cheng's (1997) power PC theory attempts to make these assumptions explicit, providing an axiomatic characterization of causality and proposing that human causal judgments correspond to "causal power," the probability that C produces E in the absence of all other causes. Causal power for a generative cause can be estimated from contingency data, with Cheng (1997) giving the expression:

$$\text{power} = \frac{\Delta P}{1 - P(e^+|c^-)}.$$

(2)

Causal power takes ΔP as a component, but predicts that ΔP will have a greater effect when $P(e^+|c^-)$ is large. Causal power can also be evaluated for preventive causes, following from a similar set of assumptions about the nature of such causes. The causal power for a preventive cause is

$$\text{power} = -\frac{\Delta P}{P(e^+|c^-)}.$$

(3)

For preventive causes, the effect of $P(e^+|c^+)$ on causal power is reversed, with ΔP having a greater influence when $P(e^+|c^+)$ is small.

The measure of causal power in Eq. (2) can be derived from a counterfactual treatment of "sufficient cause" (Pearl, 2000). Causal power corresponds to the probability that, for a case in which C was not present and E did not occur, E would occur if C was introduced. This probability depends upon ΔP,

corresponding to the raw increase in occurrences of E, but has to be normalized by the proportion of the cases in which C could actually have influenced E. If some of the cases already show the effect, then C had no opportunity to influence those cases and they should not be taken into account when evaluating the strength of C. The requirement of normalization introduces $P(e^-|c^-) = 1 - P(e^+|c^-)$ in the denominator. Pearl (2000) derives ΔP from a similar treatment of "necessary and sufficient cause."

To illustrate the difference between causal power and ΔP, consider the problem of establishing whether injecting chemicals into mice results in gene expression. Two groups of 60 mice are used in two experiments evaluating the effect of different chemicals on different genes. In each experiment, one group is injected with the chemical, and the other group receives no injection. In the first experiment, 30 of the uninjected mice express the gene, $P(e^+|c^-) = 0.5$, and 36 of the injected mice express it, $P(e^+|c^+) = 0.6$. In the second experiment, 54 of the uninjected mice express the gene, $P(e^+|c^-) = 0.9$, and all 60 of the injected mice express it, $P(e^+|c^+) = 1$. In each case $\Delta P = 0.1$, but the second set of results seem to provide more evidence for a relationship between the chemical and gene expression. In particular, if we imagine that the frequency of gene expression among the uninjected mice would be reproduced exactly in the other group of mice prior to injection, it seems that the first chemical produces gene expression in only six of the thirty mice who would not have otherwise expressed the gene, while *all* of the mice not expressing the gene in the second experiment have their fates altered by the injection. This difference is reflected in causal power, which is 0.2 in the first case and 1 in the second.

2.2. The debate over rational models

ΔP and causal power make different predictions about the strength of causal relationships, and several experiments have been conducted with the aim of determining which model gives a better account of human data (e.g., Buehner & Cheng, 1997; Collins & Shanks, submitted; Lober & Shanks, 2000; Perales & Shanks, 2003; Shanks, 2002; Vallee-Tourangeau, Murphy, Drew, & Baker, 1998). Each model captures some of the trends identified in these experiments, but there are several results that are predicted by only one of the models, as well as phenomena that are predicted by neither. These negative results are almost equally distributed between the two models, and suggest that there may be some basic factor missing from both. The problem can be illustrated by considering two sets of experiments: those conducted by Buehner and Cheng (1997) and Lober and Shanks (2000).

The experiments conducted by Buehner and Cheng (1997; Buehner et al., 2003) explored how judgments of the strength of a causal relationship vary when ΔP is held constant. This was done using an experimental design adapted from Wasserman et al. (1993), giving 15 sets of contingencies expressing all possible combinations of $P(e^+|c^-)$ and ΔP in increments of 0.25. Experiments were conducted with both generative causes, for which C potentially increases the frequency of E as in the cases

described above, and preventive causes, for which C potentially decreases the frequency of E, and with both online and summary formats. For the moment, we will focus on the online study with generative causes (Buehner & Cheng, 1997, Experiment 1B), where a total of 16 trials gave the contingency information. The results of this experiment showed that at constant values of ΔP, people made judgments that were sensitive to the value of $P(e^+|c^-)$. Furthermore, this sensitivity was consistent with the role of $P(e^+|c^-)$ in causal power.

However, as was pointed out by Lober and Shanks (2000), the results also proved problematic for the Power PC theory. The design used by Buehner and Cheng (1997) provides several situations in which sets of contingencies give the same value of causal power. The data are shown in Fig. 1, together with the values of ΔP and causal power. ΔP and causal power gave r scores of .889 and .881, respectively, with scaling
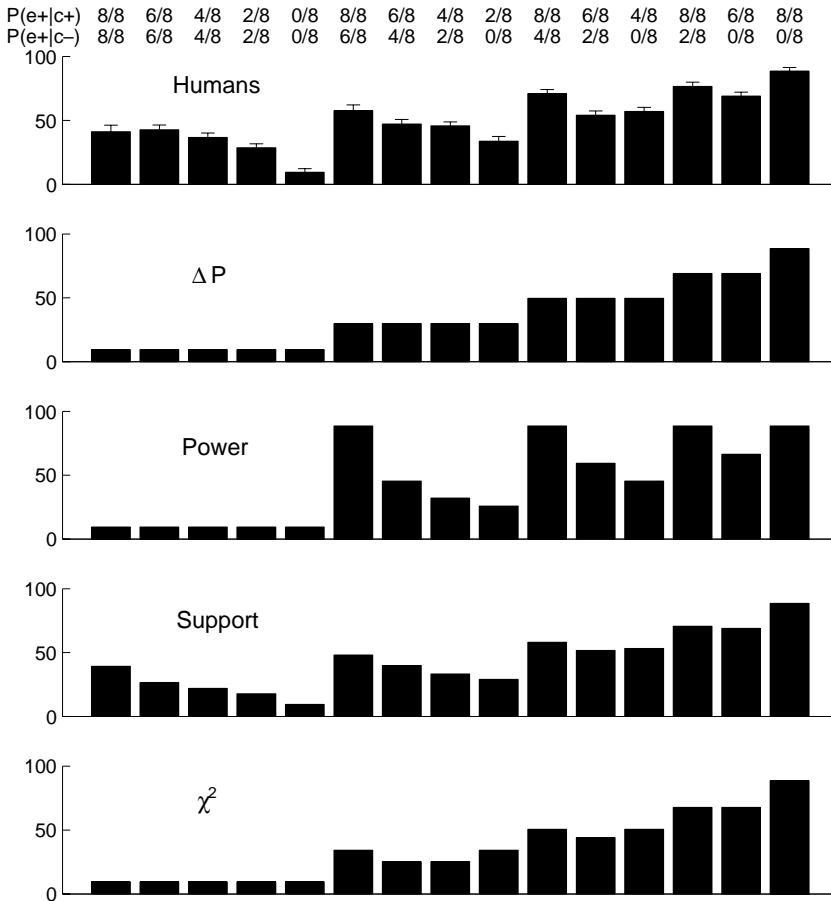

Fig. 1. Predictions of rational models compared with the performance of human participants from Buehner and Cheng (1997, Experiment 1B). Numbers along the top of the figure show stimulus contingencies, error bars indicate one standard error.

parameters $\gamma = 0.98$, 1.05.[2] As can be seen from the figure, both ΔP and causal power predict important trends in the data, but since these trends are orthogonal, neither model provides a full account of human performance. The only sets of contingencies for which the two models agree are those where ΔP is zero. For these cases, both models predict negligible judgments of the strength of the causal relationship. In contrast to these predictions, people give judgments that seem to decrease systematically as $P(e^+|c^-)$ decreases. Similar effects with $\Delta P = 0$ have been observed in other studies (e.g., Allan & Jenkins, 1983; Shanks et al., 1996), where the phenomenon is referred to as the "frequency illusion."

In a further test of the two theories, Lober and Shanks (2000) conducted a series of experiments in which either causal power or ΔP was held constant while the other varied. These experiments used both online (Experiments 1–3) and summary (Experiments 4–6) formats. The results showed systematic variation in judgments of the strength of the causal relationship at constant values of causal power, in a fashion consistent with ΔP. The results of Experiments 4–6 are shown in Fig. 2, together with the values of ΔP and causal power. The models gave r scores of .980 and .581, respectively, with $\gamma = 0.8$, 1.1. While ΔP gave a good fit to these data, the human judgments for contingencies with $\{P(e^+|c^+), P(e^+|c^-)\}$ of $\{30/30, 18/30\}$, $\{24/30, 12/30\}$, $\{12/30, 0/30\}$ are not consistent with ΔP: they show a non-monotonic trend, with smaller judgments for $\{24/30, 12/30\}$ than for either of the extreme cases. The quadratic trend over these three sets of contingencies was statistically significant, but Lober and Shanks (2000) stated that "... because the effect was non-linear, it probably should not be given undue weight" (p. 209). For the purposes of Lober and Shanks, this effect was not important because it provided no basis for discrimination between ΔP and causal power: neither of these theories can predict a non-monotonic change in causal judgments as a function of the base-rate probability $P(e^+|c^-)$.

The results of Buehner and Cheng (1997) and Lober and Shanks, 2000 illustrate that neither ΔP nor causal power provides a full account of people's judgments in causal induction tasks. These are not isolated results: ΔP and causal power cannot explain several other phenomena of human causal induction. One of these phenomena is the effect of sample size: both ΔP and causal power are defined using the conditional probabilities $P(e|c)$, and are thus insensitive to the number of observations expressing those probabilities. However, human judgments change as the number of observations contributing to a contingency table varies (White, 1998, 2002c, 2003c). Another is inferences from incomplete data: people can assess causal relationships in circumstances where there is not enough information to compute the

---

[2] In each case where we have fit a computational model to empirical data, we have used a scaling transformation to account for the possibility of non-linearities in the rating scale used by participants. This is not typical in the literature, but we feel it is necessary to separate the quantitative predictions from a dependency on the linearity of the judgment scale—an issue that arises in any numerical judgment task. We use the transformation $y = \text{sign}(x)\,\text{abs}(x)^\gamma$, where y are the transformed predictions, x the raw predictions, and γ a scaling parameter selected to maximize the linear correlation between the transformed predictions and the data. This power law transformation accommodates a range of non-linearities.
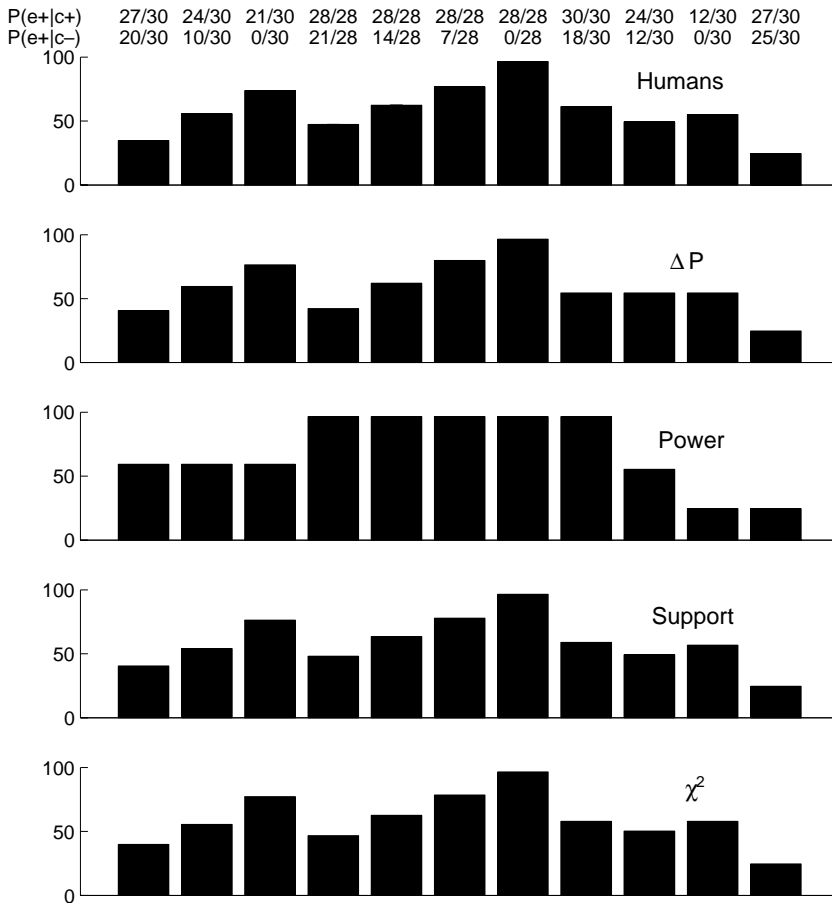
Fig. 2. Predictions of rational models compared with the performance of participants from Lober and Shanks (2000, Experiments 4–6). Numbers along the top of the figure show stimulus contingencies.

conditional probabilities of the effect in the presence and the absence of the cause. In the early stages of both everyday and scientific inferences, we might be presented with an incomplete contingency table. Neither ΔP nor causal power can explain the judgments that people make from such data. Finally, people are able to learn about causal relationships from data other than contingencies, such as the rate at which an effect is observed in the presence and absence of a cause (e.g., Anderson & Sheu, 1995). ΔP and causal power are not defined for rate data, even though it is extremely similar to contingency data.

In the remainder of the paper, we will use a computational framework based on causal graphical models to provide insight into the problems of ΔP and causal power, and to derive a new model of causal induction. We will argue that the difficulties faced by ΔP and causal power arise because people are often sensitive to the structural question of whether or not a causal relationship exists, and both ΔP and

causal power model human judgments purely in terms of the strength of a causal relationship. Our framework makes the distinction between structure and strength precise, and allows us to define a new model, causal support, which addresses this structural question. We will show that causal support accurately predicts human judgments in all of the settings mentioned above.

3. Causal graphical models

Our framework for analyzing elemental causal induction will use causal graphical models, a formalism for learning and reasoning about causal relationships that is a current topic of research in computer science and statistics (e.g., Pearl, 2000; Spirtes et al., 1993) and is beginning to be applied in cognitive science (Danks & McKenzie, submitted; Gopnik et al., 2004; Glymour, 1998, 2001; Lagnado & Sloman, 2002; Rehder, 2003; Steyvers et al., 2003; Tenenbaum & Griffiths, 2001, 2003; Waldmann & Martignon, 1998). Causal graphical models, also known as causal Bayesian networks or causal Bayes nets, provide a means of specifying the causal relationships that hold among a set of variables. Our brief summary of causal graphical models will touch on three important issues: causal structure, functional causal relationships, and the difference between learning structure and estimating parameters.

3.1. Causal structure

A graphical model provides an intuitive representation for the causal structure relating a set of variables. Nodes in the graph represent variables, and directed edges represent causal connections between those variables (Glymour, 1998, 1999; Pearl, 2000; Spirtes et al., 1993). The result is a directed graph, with "parent" nodes having arrows to their "children." For example, consider the directed graphs denoted Graph 0 and Graph 1 in Fig. 3, which we will later use in describing our framework for elemental causal induction. Both graphs are defined over three binary variables—an effect E, a potential cause C, and a background cause B, capturing the combined

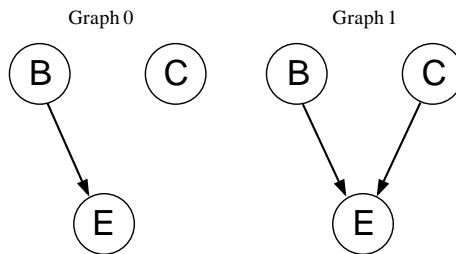Fig. 3. Directed graphs involving three variables, B, C, and E, relevant to elemental causal induction. B represents background variables, C a potential causal variable, and E the effect of interest. Graph 1, is assumed in computing ΔP and causal power. Computing causal support involves comparing the structure of Graph 1 to that of Graph 0, in which C and E are independent.

effect of all other causes of E. Each graph represents a hypothesis about the causal relations that could hold among these variables. In Graph 0, B causes E, but C has no relationship to either B or E. In Graph 1, both B and C cause E.

3.2. Functional causal relationships

The graph structure used in a causal graphical model identifies the causal relationships among variables, but says nothing about the precise nature of these relationships—they could be deterministic or probabilistic, and multiple causes of an effect could act independently or interact strongly. The functional form of a causal relationship encodes all of this information, and translates a causal structure into a probability distribution over the variables that form its nodes.

Any causal graphical model with variables $\{X_1, \ldots, X_n\}$ implies a probability distribution of the form $P(X_1, \ldots, X_n) = \prod_i P(X_i|\text{Pa}(X_i))$, where $\text{Pa}(X_i)$ is the set of parents of the node associated with X_i. This factorization of the probability distribution follows from the assumption that each variable X_i is independent of all of its non-descendants in the graph when conditioned upon its causes, $\text{Pa}(X_i)$. This assumption is called the causal Markov condition, and the relationship between statistical dependence and causation that it implies forms the basis of many algorithms for learning causal structure (e.g., Pearl, 2000; Spirtes et al., 1993).

While causal structure determines the dependencies expressed in a distribution, the actual probability distribution depends upon the choice of the conditional probabilities $P(X_i|\text{Pa}(X_i))$. Specifying these probabilities requires choosing a parameterization for each node, stating how the probability distribution over that variable depends upon the values of its parents. This parameterization determines the functional form of the causal relationship. Sometimes the parameterization is trivial—for example, in Graph 0, we need to specify only $P_0(E|B)$, where the subscript indicates that this probability is associated with Graph 0. This can be done using a single numerical parameter w_0 which provides the probability that the effect will be present in the presence of the background cause, $P_0(e^+|b^+; w_0) = w_0$. However, when a node has multiple parents, there are many different ways in which the functional relationship between causes and effects could be defined. For example, in Graph 1 we need to account for how the causes B and C interact in producing the effect E.

The conditional probability distribution associated with a node can be any probabilistically sound function of its parents, including a deterministic function. Different assumptions about the kinds of mechanisms in a domain naturally lead to different functional forms for this dependency, so there will not be a single functional form that can be used to characterize all settings in which causal learning takes place. Here, we consider three simple functional forms for the relationship between B, C, and E: noisy-OR, noisy-AND-NOT, and linear. These functional forms characterize some of the different ways in which causes combine to influence their effects, and will be used in our analysis of previous accounts of causal induction. All three parameterizations assume that E is a binary random variable—on each trial, E either occurs or does not. Later in the paper, in Section 12, we will introduce a

Poisson parameterization that can be used when the available data concern the rate at which the effect occurs in an interval.

The noisy-OR parameterization (Pearl, 1988) results from a natural set of assumptions about the relationship between cause and effect. For Graph 1, these assumptions are that B and C are both generative causes, increasing the probability of the effect; that the probability of E in the presence of just B is w_0, and in the presence of just C is w_1; and that, when both B and C are present, they have independent opportunities to produce the effect. This gives

$$P_1(e^+|b, c; w_0, w_1) = 1 - (1 - w_0)^b(1 - w_1)^c. \tag{4}$$

where w_0, w_1 are parameters associated with the strength of B, C, respectively, and $b^+ = c^+ = 1, b^- = c^- = 0$ for the purpose of arithmetic operations. This expression gives w_0 for the probability of E in the presence of B alone, and $w_0 + w_1 - w_0 w_1$ for the probability of E in the presence of both B and C. This parameterization is called a noisy-OR because if w_0 and w_1 are both 1, Eq. (4) reduces to the logical OR function: the effect occurs if and only if B or C are present. With w_0 and w_1 in the range $[0, 1]$ it generalizes this function to allow probabilistic causal relationships.

An analogous parameterization for preventive causes can be derived from a set of assumptions similar to those made in the noisy-OR. In the case of Graph 1, these assumptions are that B has the opportunity to produce E with probability w_0, and C independently prevents E from occurring with probability w_1. The resulting noisy-AND-NOT generalizes the logical statement that E will occur if B occurs and not C, allowing the influence of these factors to be probabilistic. The conditional probability can be written as

$$P_1(e^+|b, c; w_0, w_1) = w_0^b(1 - w_1)^c, \tag{5}$$

which gives w_0 for the probability of E in the presence of just B, and $w_0(1 - w_1)$ when both B and C are present. As with the noisy-OR, both w_0 and w_1 are constrained to lie in the range $[0,1]$.

Finally, a linear parameterization of Graph 1 assumes that the probability of E occuring is a linear function of B and C. This corresponds to assuming that the presence of a cause simply increases the probability of an effect by a constant amount, regardless of any other causes that might be present. There is no distinction between generative and preventive causes. The result is

$$P_1(e^+|b, c; w_0, w_1) = w_0 \cdot b + w_1 \cdot c. \tag{6}$$

This parameterization requires that we constrain $w_0 + w_1$ to lie between 0 and 1 to ensure that Eq. (6) results in a legal probability distribution. Because of this dependence between parameters, this parameterization is not normally used for causal graphical models. A similar set of assumptions can be captured using the more standard logistic parameterization (e.g., Neal, 1992), in which the parameters for different causes can vary independently from strongly positive (generative) to strongly negative (preventive). We consider the linear parameterization here because, as we show in the next section, it implicitly underlies one of the leading psychological models of causal judgment.

3.3. Structure learning and parameter estimation

Constructing a causal graphical model from a set of observed data involves two kinds of learning: structure learning and parameter estimation. Structure learning refers to identification of the topology of the causal graph, while parameter estimation involves determining the parameters of the functional relationships between causes and effects for a given causal structure. Structure learning is arguably more fundamental than parameter estimation, since the parameters can only be estimated once the structure is known. Learning the causal structure among many variables is a difficult computational problem, as the number of possible structures is a super-exponential function of the number of variables.

There are several approaches to parameter estimation in graphical models (e.g., Heckerman, 1998). The simplest approach is maximum-likelihood estimation. For a particular graphical structure and parameterization, the likelihood of a set of parameters given data D is the probability of D under the distribution specified by that structure, parameterization, and choice of parameter values. For example, if D is summarized by contingencies $N(e,c)$, the likelihood for the parameters w_0, w_1 in Graph 1 is given by

$$P_1(D|w_0, w_1, \text{Graph 1}) = \prod_{e,c} P_1(e|c, b^+; w_0, w_1)^{N(e,c)}, \tag{7}$$

where the product is taken over all pairs of e^+, e^- and c^+, c^-. A maximum-likelihood estimate is a choice of values for w_0, w_1 that maximizes $P_1(D|w_0, w_1, \text{Graph 1})$.

Structure learning attempts to identify the causal structure underlying a set of observed data. There are two major approaches to structure learning: constraint-based learning, and Bayesian inference. In constraint-based learning, constraints on possible causal structures are inferred by evaluating the statistical dependencies among variables, and sets of causal structures that satisfy these constraints are logically derived from these constraints (e.g., Pearl, 2000; Spirtes et al., 1993). In contrast, the Bayesian approach to structure learning evaluates each causal structure in terms of the probability it assigns to a dataset. By integrating over the values that parameters could assume, it is possible to compute the probability of a dataset given a graphical structure without committing to a particular choice of parameter values (e.g., Cooper & Herskovits, 1992). This computation provides $P(D|\text{Graph } i)$, and a posterior probability distribution over graphs, $P(\text{Graph } i|D)$ can be obtained by applying Bayes' rule:

$$P(\text{Graph } i|D) \propto P(D|\text{Graph } i)P(\text{Graph } i), \tag{8}$$

where $P(\text{Graph } i)$ is a prior probability distribution over graphs. Often, priors are either uniform (giving equal probability to all graphs), or give lower probability to more complex structures. Bayesian structure learning proceeds by either searching the space of structures to find that with the highest posterior probability (Friedman, 1997; Heckerman, 1998), or evaluating particular causal relationships by integrating over the posterior distribution over graphs (Friedman & Koller, 2000).

4. A framework for elemental causal induction

Framing the problem of causal induction in terms of causal graphical models reveals that it has two components: structure learning and parameter estimation. Given an always-present background variable B, a potential cause C, and a potential effect E, together with contingency information about the co-occurrence of C and E, the structure learning component is the assessment of whether or not a causal relationship in fact exists between C and E. This can be formalized as a decision between the structures Graph 1 and Graph 0 in Fig. 3. The parameter estimation component is the assessment of the strength of this relationship: assuming Graph 1 is appropriate, and determining the value of the parameter describing the relationship between C and E.

The distinction between causal structure and causal strength is important in scientific inference. When scientists investigate a phenomenon, they are typically interested in two questions: whether an effect exists, and how strong this effect might be. Different statistical tools are used for answering these questions. In the frequentist approach to statistics commonly used in psychological experimentation, statistical hypothesis testing is applied to the former, and measures of effect size to the latter. Statistical significance is a function of both effect size and sample size. As a consequence, there is no logically necessary relationship between causal structure and apparent causal strength: small effects can be statistically significant, and large effects in small samples may not be supported by sufficient data to guarantee rejection of a null hypothesis.

In this section, we will show that the distinction between structure and strength can be used to shed light on rational models of human causal induction. Both ΔP and causal power address only the parameter estimation component of elemental causal induction, providing measures of the strength of a causal relationship. We will describe a rational model that addresses the structure learning component, which we term "causal support."

4.1. Causal induction as parameter estimation: ΔP and causal power

The two rational models at the heart of the current theoretical debate about elemental causal induction address the same component of the underlying computational problem in fundamentally similar ways. Both ΔP and causal power are maximum-likelihood estimates of the causal strength parameter w_1 in Graph 1, but under different parameterizations (Tenenbaum & Griffiths, 2001). As shown in Appendix A, ΔP corresponds to the linear parameterization (Eq. (6)), whereas causal power for generative causes corresponds to the noisy-OR parameterization (Eq. (4)) and for preventive causes corresponds to the noisy-AND-NOT parameterization (Eq. (5)). Glymour (1998) also showed that causal power corresponds to a direct estimate for the strength parameter in a noisy-OR. Since they are estimates of the parameters of a fixed graphical structure, ΔP and causal power both measure the strength of a causal relationship, based upon the assumption that the relationship exists.

As point estimates of a parameter, ΔP and causal power share several properties. Firstly, they do not answer the question of whether or not a causal relationship exists. Large values of ΔP or causal power are likely to be associated with genuine causal relationships, but small values are non-diagnostic. Shanks (1995a, p. 260) notes this, pointing out that while ΔP for the effect of smoking on lung cancer is small, around 0.00083, "no one would deny that the relationship between smoking and lung cancer is an important one." This relates to a second property of both ΔP and causal power: these measures contain no information about uncertainty in the estimates involved. This uncertainty is crucial to deciding whether a causal relationship actually exists. Even a small effect can provide strong evidence for a causal relationship, provided its value is estimated with high certainty. The insensitivity to sample size exhibited by ΔP and causal power is one consequence of not incorporating uncertainty.

Identifying both ΔP and causal power as maximum-likelihood parameter estimates also helps to illustrate how these measures differ: they make different assumptions about the functional form of a causal relationship. The linear relationship assumed by ΔP seems less consistent with the intuitions people express about causality than the noisy-OR, an important insight which is embodied in Cheng's (1997) power PC theory. Cheng's (1997) distinction between "causal" and "covariational" measures turns on this fact: she views the noisy-OR parameterization as resulting from the correct set of assumptions about the nature of causality, and it is the use of this parameterization that distinguishes the Power PC theory from covariation-based accounts. We suspect that the appropriate parameterization for the relationship between a cause and its effects will depend upon an individual's beliefs about the causal mechanism by which those effects are brought about. For some causal mechanisms, other parameterizations may be more appropriate than the noisy-OR.

4.2. Causal induction as structure learning: causal support

In terms of the graphical models in Fig. 3, ΔP and causal power are both concerned with the problem of estimating the parameters of Graph 1. However, much of human causal induction seems to be directed at the problem of inferring the qualitative causal structure responsible for a set of observations. In the case of elemental causal induction, this problem reduces to deciding whether a set of observations were generated by Graph 0, in which C does not cause E, or Graph 1, in which C causes E. This binary decision is a result of the deterministic nature of the existence of causal relationships—either a relationship exists or it does not—a property of causality that is not captured by considering only the strength of a causal relationship (c.f. Goldvarg & Johnson-Laird, 2001).

The structural inference as to whether the data were generated by Graph 0 or Graph 1 can be formalized as a Bayesian decision. Making this decision requires evaluating the evidence the data provide for a causal relationship between C and E—determining the extent to which those data are better accounted for by Graph 1 than Graph 0. Having specified two clear hypotheses about the source of a set

of contingency data D, we can then compute the posterior probabilities of Graphs 0 and 1 by applying Bayes' rule. The posterior probability of Graph 1 indicates the extent to which a rational learner should believe in the existence of a causal relationship after observing D, but it may be more appropriate to model human judgments using a directly comparative measure, such as the log posterior odds (c.f. Anderson, 1990; Shiffrin & Steyvers, 1997). In log odds form, Bayes' rule is

$$\log \frac{P(\text{Graph }1|D)}{P(\text{Graph }0|D)} = \log \frac{P(D|\text{Graph }1)}{P(D|\text{Graph }0)} + \log \frac{P(\text{Graph }1)}{P(\text{Graph }0)}, \tag{9}$$

where the left-hand side of the equation is the log posterior odds, and the first and second terms on the right-hand side are the log likelihood ratio and the log prior odds, respectively.

Bayes' rule stipulates how a rational learner should update his or her beliefs given new evidence, with the log posterior odds combining prior beliefs with the implications of the evidence. The effect of data D on the belief in the existence of a causal relationship is completely determined by the value of the log likelihood ratio. The log likelihood ratio is thus commonly used as a measure of the evidence that data provide for a hypothesis, and is also known as a Bayes factor (Kass & Rafferty, 1995). We will use this measure to define "causal support," the evidence that data D provide in favor of Graph 1 over Graph 0

$$\text{support} = \log \frac{P(D|\text{Graph }1)}{P(D|\text{Graph }0)}. \tag{10}$$

To evaluate causal support, it is necessary to compute $P(D|\text{Graph }1)$ and $P(D|\text{Graph }0)$. Using a procedure described in Appendix A, it is possible to compute these probabilities without committing to particular values of the parameters w_0 and w_1 by integrating over all possible values these parameters could assume.

In computing causal support for binary-valued effects, we use the noisy-OR parameterization (Eq. (4)) for generative causes, and the noisy-AND-NOT parameterization (Eq. (5)) for preventive causes. This choice of parameterizations seems appropriate for capturing the assumptions behind problems like evaluating the influence of chemicals on gene expression, where each cause should have an independent opportunity to influence the effect. Since causal power is a maximum-likelihood estimator of w_1 under this parameterization, this results in a relationship between causal support and causal power. Speaking loosely, causal support is the Bayesian hypothesis test for which causal power is an effect size measure: it evaluates whether causal power is significantly different from zero. Causal support can also be defined for models with different functional dependencies and with different kinds of variables, a fact that we will exploit in Section 12 when we consider causal learning from rates.

As illustrated in Fig. 4, the major determinant of causal support is the extent to which the posterior distribution over w_1 places its mass away from zero. The contingency data for the top three cases shown in the figure all result in the same estimate of causal power (approximately the peak of the posterior distribution on w_1), but increasing the number of observations contributing to these contingencies decreases
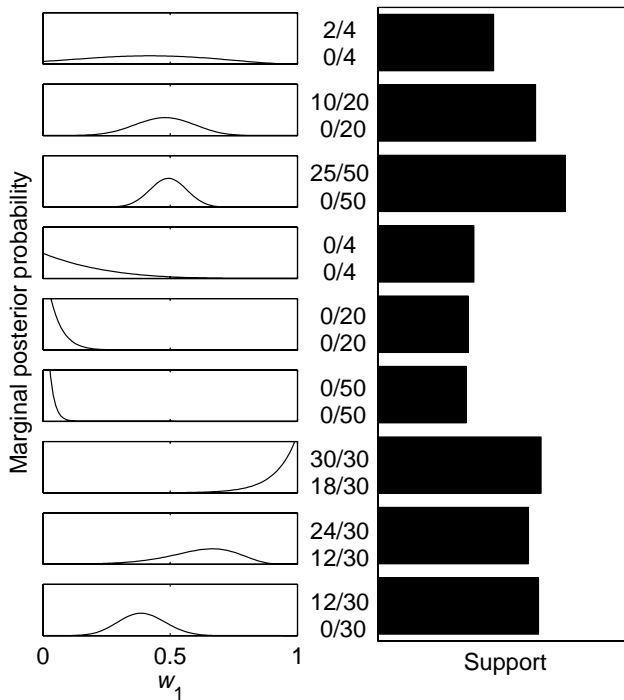
Fig. 4. Marginal posterior distributions on w_1 and values of causal support for six different sets of contingencies. The first three sets of contingencies result in the same estimates of ΔP and causal power, but different values of causal support. The change in causal support is due to the increase in sample size, which reduces uncertainty about the value of w_1. As it becomes clear that w_1 takes on a value other than zero, the evidence for Graph 1 increases, indicated by the increase in causal support. The second set of three contingencies shows that increasing sample size does not always result in increased causal support, with greater certainty that w_1 is zero producing a mild decrease in causal support. The third set of three contingencies illustrates how causal support and causal power can differ. While the peak of the distribution over w_1, which will be close to the value of causal power, decreases across the three examples, causal support changes in a non-monotonic fashion.

uncertainty about the value of w_1. It thus becomes more apparent that w_1 has a value greater than zero, and causal support increases. However, higher certainty does not always result in an increase in causal support, as shown by the next three cases in the figure. Causal power is zero for all three cases, and once again the posterior distribution shows higher certainty when the number of observations is large. Greater confidence that w_1 should be zero now results in a *decrease* in causal support, although the effect is weaker than in the previous case. The last three cases illustrate how causal support can differ from causal power. The contingencies {30/30,18/30} suggest a high value for w_1, with relatively high certainty, and consequently strong causal support; {24/30,12/30} suggest a lower value of w_1, with less certainty, and less causal support; and {12/30, 0/30} produces an even lower value of w_1, but the higher certainty that this value is greater than zero results in more causal support.

4.3. Approximating causal support with χ^2

The standard frequentist analysis of contingency tables is Pearson's (1904/1948) χ^2 test for independence, or the related likelihood ratio test G^2 (e.g., Wickens, 1989). The use of the χ^2 test as a model for human causal judgment was suggested in the psychological literature (e.g., Allan, 1980), but was rejected on the grounds that it neglects the kind of asymmetry that is inherent in causal relationships, providing information solely about the existence of statistical dependency between the two variables (Allan, 1980; López et al., 1998; Shanks, 1995b). χ^2 also makes no commitment about the functional form of the relationship between cause and effect: it simply detects any kind of statistical dependency between C and E. These weak assumptions are important for χ^2 to be useful as a statistical test across a wide range of settings, but they produce problems for it as a model of human causal judgments.

As a measure of the evidence for a particular dependency structure, χ^2 is related to causal support. In Appendix A, we show that under certain conditions causal support can be approximated by Pearson's χ^2 test for independence. This approximation only holds when the contingency table contains a large number of observations, and the potential cause has a weak effect. χ^2 should be treated with caution as a model of causal induction, for exactly the reasons identified above: it is a symmetric test of statistical dependency, while causal support postulates a specific form and direction for a causal relationship. The different assumptions behind these two approaches can result in quite different predictions.

4.4. Summary

Causal induction involves two problems: evaluating whether or not a causal relationship exists, and establishing the strength of that relationship. Causal graphical models provide the tools for treating both of these problems formally, in terms of structure learning and parameter estimation. In the remainder of the paper, we will argue that the focus on parameter estimation in previous models of elemental causal induction is responsible for the inability of these models to account for several trends in human judgments, and we will show that causal support fares better in explaining these phenomena. We will examine in detail five phenomena that are problematic for existing models, but are predicted by causal support. These phenomena were introduced in Section 2.2: the interaction between ΔP and the base-rate probability of the effect, $P(e^+|c^-)$, as demonstrated in Buehner and Cheng (1997); non-monotonic judgments as a result of changes in base-rate, as seen by Lober and Shanks (2000); effects of sample size (e.g., White, 1998, 2002c, 2003c); inferences from contingency tables with empty cells; and causal induction from rates (cf. Wasserman, 1990; Anderson & Sheu, 1995).

5. Interaction between ΔP and $P(e^+|c^-)$

In Section 2.2, we discussed results from Experiment 1B of Buehner and Cheng (1997; later published in Buehner et al., 2003) which are shown in Fig. 1. This

experiment used an online format, and produced trends that were independently problematic for both ΔP and causal power, as well as a trend that neither model could predict: people's judgments at $\Delta P = 0$ decrease as the base-rate probability of the effect, $P(e^+|c^-)$, decreases. The fundamental problem in explaining the results of Buehner and Cheng (1997) is accounting for the interaction between ΔP and the base-rate probability in producing human judgments. ΔP predicts no interaction, and causal power predicts the simple relationship given in Eq. (2), but neither of these predictions matches human judgments. We will show that causal support is able to correctly predict this interaction, demonstrating that the model can capture the trends found by Buehner and Cheng (1997).

Fig. 1 shows the data from Buehner and Cheng (1997, Experiment 1B) together with predictions of four models: ΔP, causal power, causal support, and χ^2. As noted in Section 2.2, both ΔP and causal power capture some trends in the data, but miss others, resulting in correlations of $r = .889$ and $r = .881$. Causal support provides the best quantitative account of this dataset, $r = .968$, $\gamma = .668$, and accounts for all of the major trends in human judgments, including those at $\Delta P = 0$. Due to the small samples, χ^2 gives a poor approximation to causal support, and a correlation of $r = .889$, $\gamma = .596$.

Causal support correctly predicts the interaction between ΔP and $P(e^+|c^-)$ in influencing people's judgments. In particular, it is the only model that predicts human judgments at $\Delta P = 0$. The explanation for these predictions is not that there is decreasing evidence *for* a causal relationship as $P(e^+|c^-)$ decreases, but rather that there is little evidence for or against a causal relationship when $P(e^+|c^-) = 1$, and increasing evidence *against* a causal relationship as $P(e^+|c^-)$ decreases. This account depends on the assumption that the causal relationship—if it exists—is generative (increasing the probability of the effect, rather than preventing it). At one extreme, when $\{P(e^+|c^+), P(e^+|c^-)\} = \{8/8, 8/8\}$, all mice expressed the gene irrespective of treatment, and intuitively there should be no evidence for a causal relationship. But there can also be no evidence against a (generative) causal relationship, because of a complete "ceiling" effect: it is impossible for the cause to increase the probability of E occurring above its baseline value when $P(e^+|c^-) = 1$. This uncertainty in causal judgment when $P(e^+|c^-) = 1$ and $\Delta P = 0$ is predicted by both causal support, which is near 0, and also (as Cheng, 1997, points out) by causal power, which is undefined there.

Only causal support, however, predicts the gradient of judgments as $P(e^+|c^-)$ decreases. Causal support becomes increasingly negative as the ceiling effect weakens and the observation that $\Delta P = 0$ provides increasing evidence against a generative causal relationship. At the other extreme, when $P(e^+|c^-) = 0/8$, no untreated mice expressed the gene, and there are eight opportunities for a causal relationship to manifest itself in the treated mice if such a relationship in fact exists. The fact that the effect does not appear in any treated mice, $P(e^+|c^+) = 0/8$, strongly suggests that the drug does *not* cause gene expression. The intermediate cases provide intermediate evidence against a causal relationship. The contingencies $\{2/8, 2/8\}$ offer six chances for the treatment to have an effect, and the fact that it never does so is slightly weaker evidence against a relationship than in the $\{0/8, 0/8\}$ case, but more compelling than

for $\{6/8, 6/8\}$, where the cause only has two chances to manifest itself and the observation that $\Delta P = 0$ could easily be a coincidence. This gradient of uncertainty shapes the Bayesian structural inference underlying causal support, but it does not impact the maximum-likelihood parameter estimates underlying causal power or ΔP.

Fig. 5 reveals why causal support is sensitive to the change in $P(e^+|c^-)$, showing the posterior distribution on w_1 for each set of contingencies. Greater certainty in the
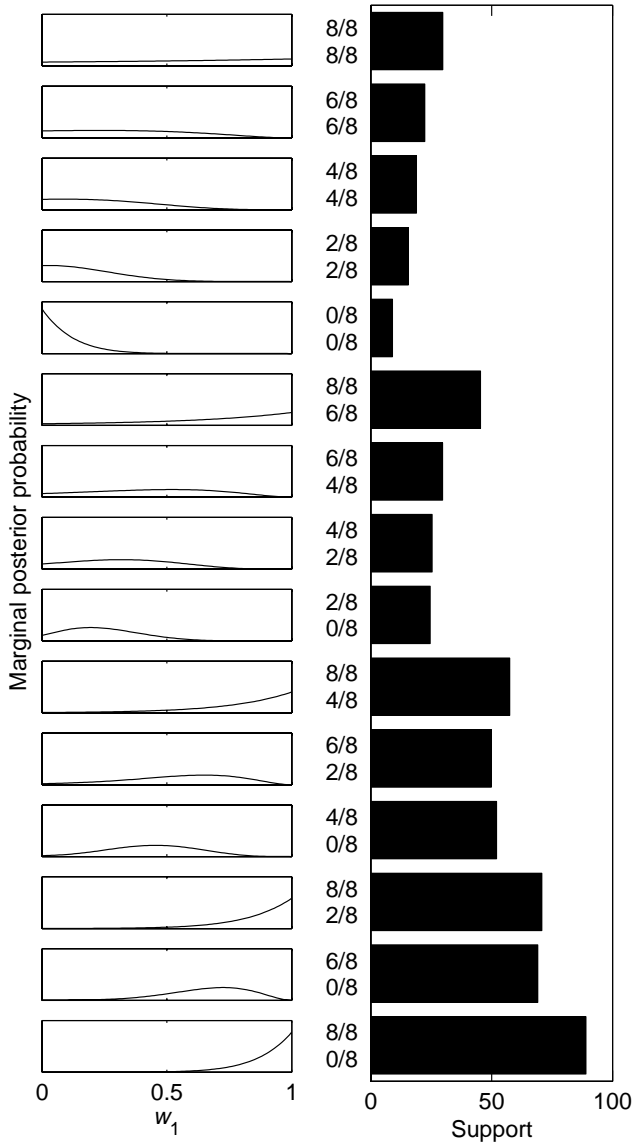

Fig. 5. Marginal posterior distributions on w_1 and values of causal support for the contingencies used in Buehner and Cheng (1997, Experiment 1B).

value of w_1 is reflected in a more peaked distribution, and causal support becomes larger as it becomes more apparent that w_1 is greater than zero. The top five plots show the cases where $\Delta P = 0$. Despite the fact that ΔP is the same in these five cases, the posterior distributions on w_1 look quite different. Four of the distributions have a maximum at $w_1 = 0$, consistent with the estimate of causal power for these contingencies, but they differ in the certainty of the estimate, reflected in the breadth of the posterior about this point.[3] As $P(e^+|c^-)$ increases, fewer observations contribute to the estimate of w_1, and the posterior becomes more diffuse, being almost uniform for $P(e^+|c^-) = 1$.

This explanation can also be applied to a similar trend that emerges with preventive causes. The results of Experiment 1A of Buehner and Cheng (1977; also published in Buehner et al., 2003), are shown in Fig. 6. This experiment followed a design analogous to Experiment 1B, but with preventive causes. Here, people's judgments for $\Delta P = 0$ decrease as $P(e^+|c^-)$ increases. Causal support, parameterized with a noisy-AND-NOT (Eq. (5)) rather than a noisy-OR, provides the best account of these data, including the trend at $\Delta P = 0$, with $r = .922$, $\gamma = 0.537$. Causal power performs similarly, $r = .912$, $\gamma = 1.278$, while ΔP gives $r = .800$, $\gamma = 0.943$ and χ^2 gives $r = .790$, $\gamma = 0.566$. The explanation for the predictions of causal power is similar to the generative case, although now the "ceiling effect" is a "floor effect": as $P(e^+|c^-)$ increases there is more opportunity for a non-zero value of w_1 to be demonstrated.

6. Non-monotonic effects of $P(e^+|c^-)$

Accounting for the interaction between ΔP and the base-rate probability, $P(e^+|c^-)$, is fundamental to explaining the results of Buehner and Cheng (1997). It is also important in explaining other phenomena of causal induction. The second dataset discussed in Section 2.2 was Experiments 4–6 from Lober and Shanks (2000), shown in Fig. 2. ΔP accounts for these data quite well, reflected in the high correlation coefficient, $r = .980$, $\gamma = 0.797$, while causal power does poorly, $r = .581$, $\gamma = 1.157$. However, neither of these models can predict the non-monotonic effect of $P(e^+|c^-)$ seen with the $\{P(e^+|c^+), P(e^+, c^-)\}$ pairs $\{30/30, 18/30\}$, $\{24/30, 12/30\}$, $\{12/30, 0/30\}$. A similar, but weaker, trend can be seen in the online data of Buehner and Cheng (1997, Experiment 1B), shown in Fig. 1, for the contingencies $\{8/8, 4/8\}, \{6/8, 2/8\}, \{4/8, 0/8\}$. These non-monotonic trends cannot even be predicted by models that form linear combinations of the entries in a contingency table, such as those of Anderson and Sheu (1995) and Schustack and Sternberg (1981), despite their many free parameters and great flexibility.

---

[3] The distribution for w_1 when $P(e^+|c^+) = P(e^+|c^-) = 1$ is mostly flat but has a very slight peak at $w_1 = 1$, despite causal power being undefined for this case. This is because there is also uncertainty in the value of w_0. If w_0 actually takes on any value less than 1, and the large number of occurrences of the effect in the absence of the cause is just a coincidence, then the large number of occurrences of the effect in the presence of the cause still needs to be explained, and the best explanation is that w_1 is high.
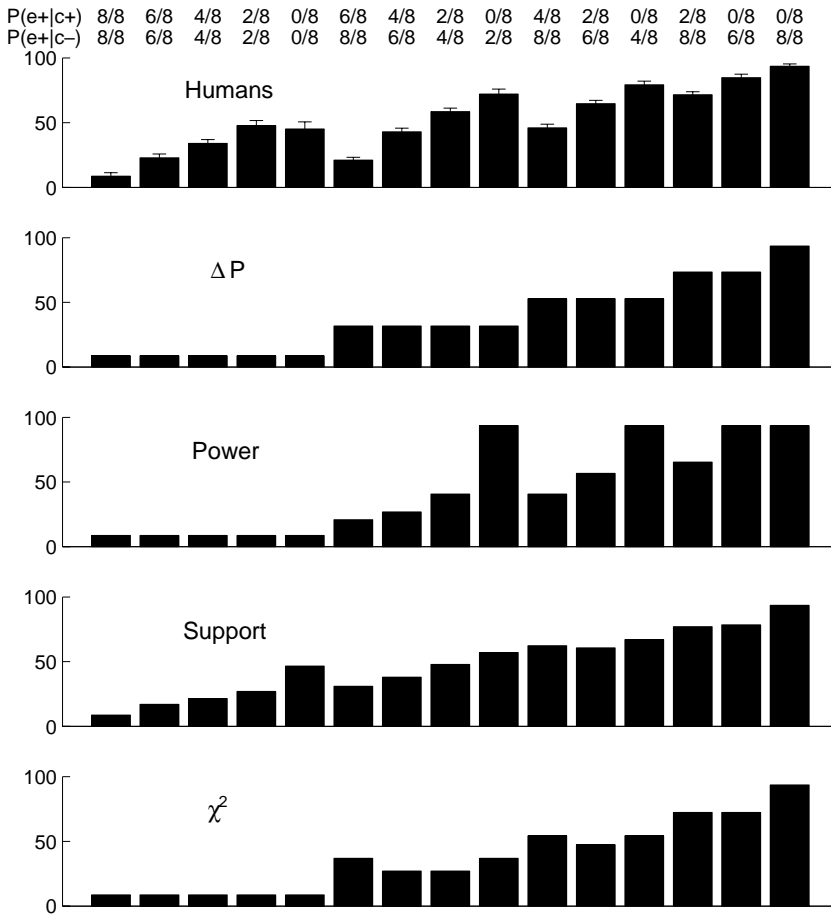
Fig. 6. Predictions of rational models compared with the performance of human participants from Buehner and Cheng (1997, Experiment 1A). Numbers along the top of the figure show stimulus contingencies, error bars indicate one standard error.

Causal support gives the best quantitative fit to this dataset, $r = .994$, $\gamma = 0.445$, with χ^2 performing similarly, $r = .993$, $\gamma = 0.502$. Both causal support and χ^2 predict non-monotonic trends, as shown in Fig. 2. The intuitive reasons for these predictions were mentioned when discussing Fig. 4, which uses exactly the same set of contingencies: while $\{24/30, 12/30\}$ suggests a higher value of causal power than $\{12/30, 0/30\}$, such a difference in contingencies is more likely to arise by chance. As with the explanation of predictions for $\Delta P = 0$ given in Section 5, the high certainty in the value of w_1 for $\{12/30, 0/30\}$ results partly from the low value of $P(e^+|c^-)$.

The non-monotonic trend observed in Experiments 4–6 of Lober and Shanks (2000) did not appear in their Experiments 1–3, despite the use of the same contingencies, as shown in Fig. 7. The only difference between these two sets of experiments
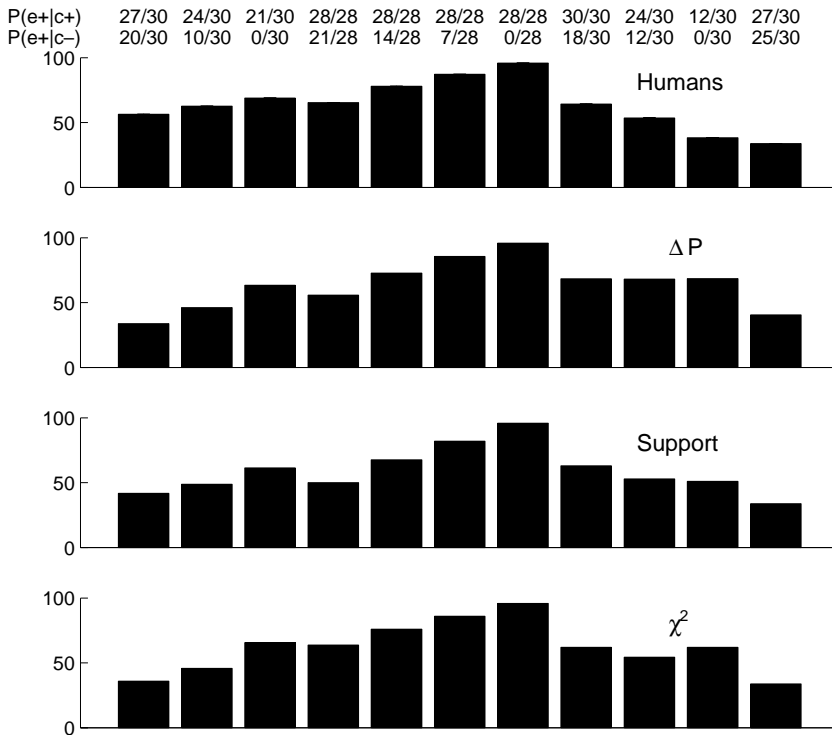
Fig. 7. Predictions of rational models compared with the performance of participants from Lober and Shanks (2000, Experiments 1–3). Numbers along the top of the figure show stimulus contingencies, but the results are constructed by averaging over the blocks of trials seen by individual subjects, in which contingencies varied.

was the presentation format, with an online format being used in Experiments 1–3, and a summary format in Experiments 4–6. This presents a challenge for the explanation based on causal support given above. However, we will argue that this discrepancy can be resolved through a finer-grained analysis of these experiments.

Catena, Maldonado, and Cándido (1998) and Collins and Shanks (2002) both found that people's judgments in online experiments are influenced by response frequency. Specifically, people seem to make judgments that are based upon the information presented since their last judgment, meaning that "primacy" or "recency" effects can be produced by using different response schedules (Collins & Shanks, 2002). Lober and Shanks (2000) used a procedure in which participants made judgments after blocks of trials, with six blocks of length 10 in Experiment 1, two blocks of length 18, and one of length 20 in Experiment 2, and 3 blocks of length 20 in Experiment 3. The actual trials presented in each block were selected at random. Thus, while the overall contingencies might match those used in Experiments 4–6, the contingencies contributing to any individual judgment varied. This may account for the difference between the results of the two experiments. In particular, the smaller sample sizes contributing to the contingencies may affect people's judgments.

Table 2
Correlations of rational models with results from Lober and Shanks (2000)

	ΔP	Support	χ^2
Experiment 1	0.354	0.350	**0.354**
Experiment 2	0.462	0.465	**0.471**
Experiment 3	0.336	**0.382**	0.303
Overall means	0.695	**0.895**	0.829

Note. Boldface indicates highest correlation in each row.

To test this hypothesis, we used the records of the individual trials seen by the participants in these experiments to establish the contingencies each participant saw in each block, and evaluated the performance of different models in predicting the judgments of individual participants for each block from these contingencies.[4] The results of these analyses are given in Table 2. The correlations are lower than those reported elsewhere in the paper because they concern the responses of individual participants rather than average scores. While all models did equally well in predicting the results of Experiments 1 and 2, causal support gave a better account of the results of Experiment 3, with a correlation of $r = .382$ as compared to $r = .336$ for ΔP, and $r = .303$ for χ^2. The model predictions, averaged across blocks and participants in the same fashion as the data, are shown in Fig. 7. The mean values of causal support do not show any predicted non-monotonicity, in accord with the data. As shown in Table 2, the mean causal support correlates better with the mean human judgments than the mean predictions of any other model, with $r = .895$ for causal support, $r = .695$ for ΔP, and $r = .829$ for χ^2.

There are two reasons why causal support does not predict a non-monotonic trend for Experiments 1–3: smaller samples, and variation in observed contingencies. The effect of sample size is simple: causal support predicts a non-monotonic trend for contingencies derived from 30 trials in each condition, but this trend is almost completely gone when there are only 10 trials in each condition. Small samples provide weaker evidence that the strength of the cause is different from zero in all conditions, with $\{12/30, 0/30\}$ and $\{24/30, 12/30\}$ giving equivalently weak support for a causal relationship. The effect of variation in observed contingencies is more complex. Since both ΔP and causal power are estimated from the conditional probabilities $P(e|c)$, and the empirical probabilities give unbiased estimates of the true probabilities, averaging across many sets of contingencies generated according to those probabilities gives mean values that approximate the true ΔP and causal power. Causal support is a more complex function of observed contingencies, and averaging causal support across a set of samples produces different results from computing causal support from the average contingencies. In particular, variation in the contingencies within

---

[4] The raw data from Lober and Shanks (2000) were supplied by Klaus Melcher. The models compared in this section were fit using the same scaling parameter for all participants within the same experiment. Causal power was not computed for this comparison, as the presence of extreme contingencies in several cases resulted in undefined values of causal power, interfering with correlations and averaging.

blocks results in some blocks providing very weak evidence for a causal relationship (for example, in the $\{12/30, 0/30\}$ condition, one participant saw a block in which the cause was present on eight trials, but the effect occurred on only one of these trials). Such results have the greatest effect in the $\{12/30, 0/30\}$ condition, and the mean causal support in this condition consequently ends up slightly lower than causal support estimated from mean contingencies.

Although a non-monotonic effect of $P(e^+|c^-)$ appears in both Lober and Shanks (2000, Experiments 4–6) and Buehner and Cheng (1997), it has not been the principal target of any experiments. Since no existing model of elemental causal induction can predict this phenomenon, confirming its existence would provide strong evidence in favor of causal support. Experiment 1 was designed to explore this non-monotonic effect further.

7. Experiment 1

7.1. Method

7.1.1. Participants
One hundred and eight Stanford undergraduates participated for course credit.

7.1.2. Stimuli
The contingencies used in the experiment are shown in Fig. 8. They included three sets of three contingencies at fixed values of ΔP but different values of $P(e^+|c^-)$, and several distractors. The sets of contingencies with fixed ΔP used $\Delta P = 0.40$, $\Delta P = 0.07$ and $\Delta P = 0.02$. ΔP predicts no effect of $P(e^+|c^-)$ within these sets, so any effect provides evidence against this model. Causal power predicts a monotonic increase in people's judgments as $P(e^+|c^-)$ increases, and causal support predicts a non-monotonic trend in the first two sets of contingencies, and a monotonic increase with $P(e^+|c^-)$ in the third. Finding non-monotonic effects in the first two sets of contingencies would thus provide evidence for causal support over causal power.

7.1.3. Procedure
The experiment was conducted in survey form. The instructions placed the problem of causal induction in a medical context:

> Imagine that you are working in a laboratory and you want to find out whether certain chemicals cause certain genes to be expressed in mice. Below, you can see laboratory records for a number of studies. In each study, a sample of mice were injected with a certain chemical and later examined to see if they expressed a particular gene. Each study investigated the effects of a *different* chemical on a *different* gene, so the results from different studies bear no relation to each other.
> Of course, these genes may sometimes be expressed in animals not injected with a chemical substance. Thus, a sample of mice who were not injected with any chemical were also checked to see if they expressed the same genes as the
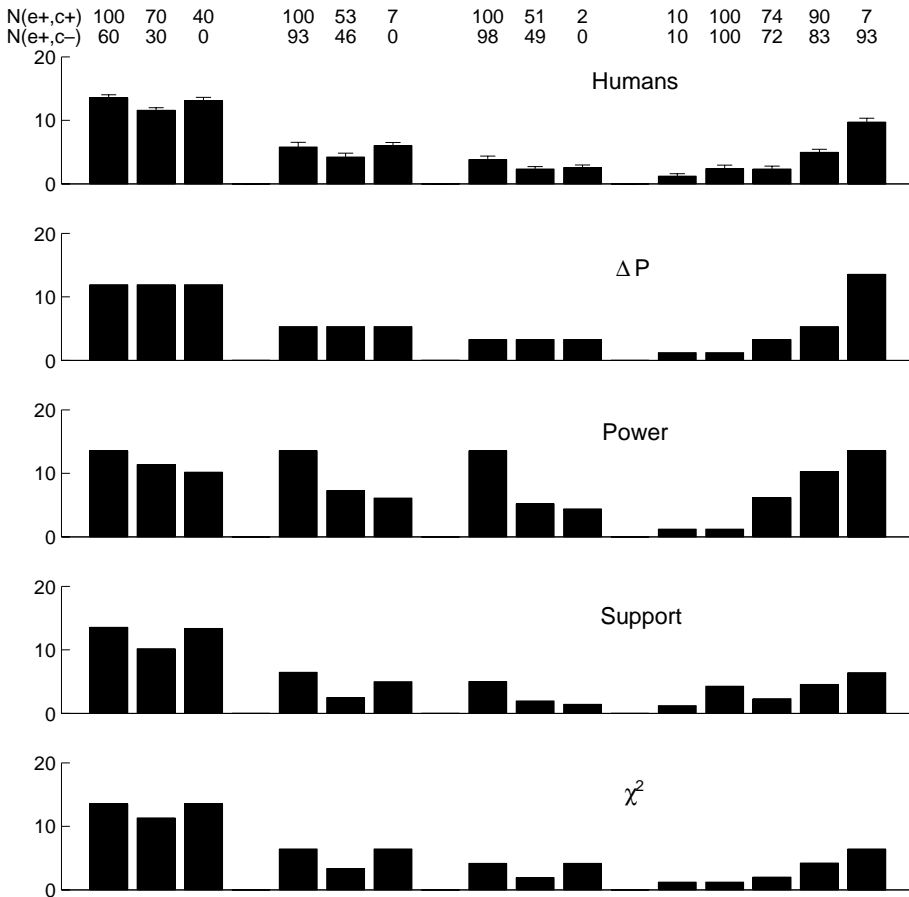
Fig. 8. Predictions of rational models compared with results of Experiment 1. Numbers along the top of the figure show stimulus contingencies. These numbers give the number of times the effect was present out of 100 trials, for all except the last column, where the cause was present on 7 trials and absent on 193. The first three groups of contingencies are organized to display non-monotonicities in judgments, the last group contains distractor stimuli. Error bars indicate one standard error.

injected mice. Also, some chemicals may have a large effect on gene expression, some may have a small effect, and others, no effect.

Participants were then asked for ratings on a total of 14 different contingency structures. The instructions for producing the ratings were:

> For each study, write down a number between 0 and 20, where 0 indicates that the chemical DOES NOT CAUSE the gene to be expressed at all, and 20 indicates that the chemical DOES CAUSE the gene to be expressed every time.

Each participant completed the survey as part of a booklet of unrelated experiments.

7.2. Results and discussion

The results are shown in Fig. 8. As predicted, there was a statistically significant effect of $P(e^+|c^-)$ in all three sets of contingencies with fixed ΔP. For $\Delta P = 0.40$, $F(2,214) = 6.61$, $MSE = 17.43$, $p < .005$, for $\Delta P = 0.07$, $F(2,214) = 3.82$, $MSE = 26.93$, $p < .05$, for $\Delta P = .02$, $F(2,214) = 6.06$, $MSE = 11.27$, $p < .005$. Since a quadratic trend analysis would only test deviation from linearity, and not non-monotonicity, we evaluated the effect of $P(e^+|c^-)$ in each of these sets of contingencies by testing each pair of means with neighboring values of $P(e^+|c^-)$. The response for $\{40/100, 0/100\}$ was significantly greater than that for $\{70/100, 30/100\}$, $t(107) = 3.00$, $p < .005$, and $\{70/100, 30/100\}$ was significantly less than $\{100/100, 60/100\}$, $t(107) = 3.81$, $p < .001$, indicating a non-monotonic trend at $\Delta P = .40$. The response for $\{7/100, 0/100\}$ was significantly greater than that for $\{53/100, 46/100\}$, $t(107) = 2.13$, $p < .05$, and $\{53/100, 46/100\}$ was significantly less than $\{100/100, 93/100\}$, $t(107) = 2.70$, $p < .01$, indicating a non-monotonic trend at $\Delta P = 0.07$. At $\Delta P = .02$, $\{2/100, 0/100\}$ was greater than $\{51/100, 49/100\}$, $t(107) = 4.05$, $p < .001$, but there was no significant difference between $\{51/100, 49/100\}$ and $\{100/100, 98/100\}$, $t(107) = 0.55$, $p = .58$, providing no evidence for non-monotonicity.

The results of this experiment suggest that the non-monotonicitic trend seen by Lober and Shanks (2000) is a robust aspect of human judgments, even though it may be a small effect. Such trends can be used to assess models of causal induction. Causal support, χ^2, and ΔP gave similarly high correlations with the experimental results, with $r = .952$, $\gamma = .604$, $r = .965$, $\gamma = .446$, and $r = .943$, $\gamma = .568$, respectively. Causal power performed far worse, with $r = .660$, $\gamma = .305$. The statistically significant effects of $P(e^+|c^-)$ in the three sets of contingencies with fixed ΔP are contrary to the predictions of ΔP. Only causal support and χ^2 predicted the observed non-monotonic trends.

8. Sample size effects

In explaining the results of Lober and Shanks (2000, Experiments 1–3), we touched upon the issue of sample size. Sample size is an important factor that affects structure learning but not parameter estimation: larger samples provide better grounds for assessing the evidence for a causal relationship, but do not affect parameter estimation. Both ΔP and causal power are computed using the conditional probabilities $P(e|c)$, rather than the number of observations contributing to these probabilities. Consequently, they predict no variation in judgments as the number of trials on which the cause was present or absent is varied. In contrast, both causal support and χ^2 are sensitive to sample size.

There are two dimensions along which sample size might be varied: the ratio of the number of trials on which the cause is present or absent ($N(c^+)$ and $N(c^-)$), and the total number of trials (N). Variation of the ratio of $N(c^+)$ to $N(c^-)$ has been explored extensively by White (1998, 2002c, 2003c). These experiments revealed

several effects of sample size, inconsistent with the predictions of ΔP and causal power, and were taken to support White's (2002c) proportion of Confirming Instances (pCI) model, which differs from ΔP only when $N(c^+) \neq N(c^-)$. We will discuss the pCI model further in Section 15. We applied all five models—ΔP, causal power, pCI, causal support, and χ^2—to the results of these experiments, producing the correlations shown in Table 3. For several of these experiments, ΔP was constant for all stimuli, resulting in a correlation of 0 between ΔP and human judgments. We also present the results of the five models on Experiment 1 of Anderson and Sheu (1995), in which the entries in single cells of the contingency table were varied systematically, resulting in some sample size variation as well as other effects. The observed effects of sample size were broadly consistent with causal support, which gave either the best or close to the best account of 10 of the 12 datasets. There was no systematic relationship between the format in which contingency information was presented and the performance of the models, although causal support gave the best correlations with the three online datasets.

Variation of the total number of trials producing a set of contingencies, N, has been studied less extensively. White (2003c, Experiment 4), conducted an experiment in which this quantity was varied, and found no statistically significant effect on human ratings. As shown in Fig. 4, causal support makes clear predictions about the effect of N on the evidence for a causal relationship, and the demonstration of such an effect would provide evidence for the involvement of structure learning in human causal induction. One possible explanation for the lack of a sample size effect in White's (2003c) experiment is the use of ratings as a dependent measure: the effect of sample size might be concealed by allowing people the possibility of giving equal ratings to different stimuli. Consequently, we decided to explore this phenomenon further, using a more sensitive response measure: Experiment 2 was designed to

Table 3
Correlations of rational models with sample size experiments

Paper	Experiment	Format	ΔP	Power	pCI	Support	χ^2
White (1998)	3 (seeds)	Summary (16)	0.907	0.902	**0.929**	0.924	0.865
	3 (contentless)	Summary (16)	0.922	0.867	0.933	**0.935**	0.885
White (2002c)	1	List (8)	**0.956**	0.765	**0.956**	0.938	0.936
	2	List (12)	0.772	0.852	0.760	**0.916**	0.830
	3	List (6)	0	0.760	**0.941**	0.837	0.146
White (2003c)	1	List (8)	0.200	0.389	0.818	**0.854**	0.791
	2	Online (8)	0.070	0.409	0.706	**0.812**	0.640
	3	Summary (8)	0.392	0.383	0.467	**0.677**	0.586
	4	List (8)	0	0.037	**0.860**	0.679	0.729
	5	List (8)	0	0.373	0.788	**0.803**	0.631
	6	Online (4)	0	0	0.425	**0.676**	0
Anderson and Sheu (1995)	1	Online (80)	0.884	0.816	0.877	**0.894**	0.329

Note. Boldface indicates highest correlation in each row. Number in parentheses in format column indicates number of stimuli.

examine whether sample size affects people's assessment of causal relationships, using a rank ordering task.

9. Experiment 2

9.1. Method

9.1.1. Participants
Participants were 20 members of the MIT community who took part in the experiment in exchange for candy.

9.1.2. Stimuli
We used nine stimuli, each composed of different contingency data. The two critical sets of contingencies were a set for which $\Delta P = 0$, consisting of $\{0/4, 0/4\}$, $\{0/20, 0/20\}$, and $\{0/50, 0/50\}$, and a set for which $\Delta P = 0.5$, $\{2/4, 0/4\}$, $\{10/20, 0/20\}$, and $\{25/50, 0/50\}$. ΔP, pCI, and causal power are constant for these sets of stimuli, and any ordering should thus be equally likely. χ^2 predicts an increase in judgments with sample size for the $\Delta P = 0.5$ set, but is constant for the $\Delta P = 0$ set. Causal support predicts sample size should result in an increase in judgments with $\Delta P = 0.5$, and a decrease with $\Delta P = 0$, as shown in Fig. 4. The experiment also included three distractors, to conceal our manipulation: $\{3/4, 1/4\}$, $\{12/20, 8/20\}$, and $\{50/50, 0/50\}$.

9.1.3. Procedure
Participants read a description of an imaginary laboratory scenario, similar to that used in Experiment 1, and were shown nine cards that expressed the stimulus information described above. They were given the following instructions:

> Each of the cards in front of you summarizes the results of a different study. Look these summaries over carefully, and then place them in order from the study from which it seems LEAST LIKELY that the chemical causes the gene to be expressed, to the study in which it seems MOST LIKELY that the chemical causes the gene to be expressed.

The wording of the question in terms of likelihood followed the procedure reported by White (2003c). If participants asked whether cards could be ranked equally, they were told that they could order them randomly.

9.2. Results and discussion

Analysis of the orderings produced by the participants showed that 17 out of 20 ordered the stimuli with $\Delta P = 0.5$ by increasing sample size (binomial test, $p < .001$), while 16 out of 20 ordered the stimuli with $\Delta P = 0$ by decreasing sample size (binomial test, $p < .001$). We computed rank-order correlations with the responses of individual participants for each of the five models. We computed the rank-order correlations with the five models for each participant, averaging these correlations

to result in scores for each model.[5] Causal support and χ^2 performed equivalently, $\rho = 0.948$ and $\rho = 0.945$, respectively, followed by ΔP, $\rho = 0.905$, and causal power, $\rho = 0.859$. Causal support gave the highest correlation with the responses of eleven participants, causal power and χ^2 with four participants each, and ΔP with only one participant.

The results indicate that people are sensitive to sample size when making causal judgments. Specifically, increasing sample size increases judgments when effects are large, but decreases judgments for zero effects. Only causal support can explain this pattern of results. Sensitivity to sample size is a property of structure learning, not parameter estimation, and thus provides evidence that people approach problems of causal induction as structure learning problems.

10. Inferences from incomplete contingency tables

We began this paper by observing that everyday causal induction has several commonalities with the reasoning of early scientists. Among these commonalities is the need to make inferences from limited data. In many settings where people infer causal relationships, they do not have all of the information that is typically provided in causal induction tasks. Specifically, without a carefully designed experiment, we often do not know the frequency of the effect in the absence of the cause, leaving some of the cells in a contingency table empty. While our epigraph indicates the attention that James Currie paid to the number of patients who recovered both with and without treatment, such reporting was the exception rather than the rule prior to the development of modern experimentation. Many early medical texts, such as Jenner's (1798) famous treatise on the smallpox vaccine, consist of a description of a number of cases in which the treatment proved successful, providing only $N(e^+, c^+)$. To make an inference from such data, his readers had to use their expectations about the frequency of infections in the absence of treatment.

ΔP and causal power are both undefined when there are no trials on which the cause was absent, since $P(e^+|c^-)$ cannot be computed. This is a problem, as people readily make causal judgments under such circumstances. For example, suppose that a doctor claims to have invented a treatment that will cure a rare illness, Hopkins–Francis syndrome. He tells you that he has given this treatment to one patient with Hopkins–Francis syndrome, and after one month, all the patient's symptoms are gone. How much evidence does this provide for the treatment's effectiveness? It may provide some evidence, but not strong evidence, since we do not know how many patients would recover spontaneously in this interval.

---

[5] Correlations were averaged using the Fisher z transformation. These results include only 19 of the 20 participants, since causal support perfectly predicted the ordering given by one participant, resulting in an infinite z score. The reported mean correlation is thus an underestimate for causal support. While the other models do not predict an ordering for the two critical sets, they do predict an ordering among the full set of nine stimuli, hence $\rho > 0$.

A few months later, the doctor tells you that he has now given the treatment to three patients, and after one month all of their symptoms are gone. These data provide stronger evidence, but not that much stronger. The evidence is strengthened once more when, a few months later, the doctor tells you that he has given the treatment to twenty patients, and after one month all of their symptoms are gone. Finally, the doctor tells you that he has also seen twenty patients over the same time period who received a placebo instead of the new treatment, and all people in this group still had symptoms after a month of observation. Moreover, the people who received the treatment or the placebo were chosen at random. Now this provides very strong evidence for the treatment's effectiveness.

We can identify five stages in the accumulation of evidence in this example. The first stage is the baseline, with no information, and the contingencies {0/0,0/0}. After a single observation, we have {1/1,0/0}. Two more observations provide {3/3,0/0}, and 17 more successful treatments give {20/20,0/0}. Finally, the control condition provides {20/20,0/20}. ΔP and causal power can only be computed for this last case, where they both indicate strong evidence for a causal relationship, $\Delta P =$ power $= 1.00$. They are undefined for the other contingency tables, and thus cannot capture the weak but growing evidence these tables provide. Causal support is 0 for {0/0,0/0}, reflecting the lack of evidence for or against a causal relationship (negative values of causal support indicate evidence for Graph 0, while positive values indicate evidence for Graph 1). Causal support then gradually increases as the observations accumulate, taking values of 0.41, 0.73, and 1.29, before jumping dramatically to 23.32 for when the control condition is added. Unlike ΔP or causal power, causal support thus predicts our intuitive ordering of the strength of evidence provided by these five stimuli. In Experiment 3, we examined whether this ordering matched the judgments of naive participants.

11. Experiment 3

11.1. Method

11.1.1. Participants
Participants were 20 members of the MIT community who took part in the experiment in exchange for candy.

11.1.2. Stimuli
We used the five stimuli described above: {0/0,0/0}, {1/1,0/0}, {3/3,0/0}, {20/20,0/0}, and {20/20,0/20}.

11.1.3. Procedure
The procedure was identical to that of Experiment 2, with the stimuli being presented on cards and participants being asked to provide an ordering.

11.2. Results and discussion

Analysis of the orderings produced by the participants showed that 15 out of 20 perfectly reproduced the ordering predicted by causal support (binomial test, $p < .001$). The other five participants still showed some conformity to the predictions of causal support, with a mean correlation of $\rho = 0.51$. ΔP and causal power are undefined for all but one of these stimuli, preventing the computation of any correlations.

Causal support is the only model we have considered that is capable of capturing people's inferences from incomplete contingency tables. The ability to infer causal relationships from limited data is an extremely important part of causal induction in both everyday and scientific settings. Even today, many medical treatments are initially tested with small samples and incomplete contingency tables. Many doctors say they do not believe in a drug's effectiveness until it passes large-scale studies with appropriate controls, in part because standard statistical practice does not provide a rigorous way to evaluate treatment effectiveness with such limited data. However, the researchers who are actually coming up with and testing new treatments need to have some way of evaluating which treatments are promising and which are not, or they would never make any progress. Causal support provides an account of the rational basis of these intuitions.

12. Learning from rates

Several studies of elemental causal induction have gone beyond inferences from contingency table data, examining how people learn about causal relationships from rate data (e.g., Anderson & Sheu, 1995; Wasserman, 1990). Rates are closely related to contingencies, being the number of times the effect occurs in a continuous interval rather than the number of times the effect occurs on a set of discrete trials. Despite this relationship, previous models of causal induction such as ΔP and causal power have not been assessed using rate data. In this section, we will show how these models can be extended to allow inferences from rate data, and evaluate their predictions about human judgments. This novel setting provides an opportunity to test the generality of our framework, as the distinction between structure and strength holds whether the observed data are rates or contingencies.

Anderson and Sheu (1995, Experiment 2) conducted an experiment in which participants learned whether clicking on a flute icon caused a change in the rate of musical notes produced by the flute. They found that their results were poorly predicted by the difference in rates, defined as

$$\Delta R = N(c^+) - N(c^-), \tag{11}$$

where $N(c^+)$ is the number of events in the interval when the cause, in this case clicking on the flute, was present, and $N(c^-)$ is the number of events when the cause was absent. Anderson and Sheu (1995) found that performance could be better predicted by "grating contrast," which they defined as

$$\text{contrast} = \frac{N(c^+) - N(c^-)}{N(c^+) + N(c^-)} \tag{12}$$

and justified by its use as a measure of contrast in psychophysical research. They gave no theoretical motivation for using this measure.

In the remainder of this section, we will develop a rational account of causal induction from rates. This account makes explicit the relationship between ΔR, ΔP, and causal power, revealing that ΔR is the maximum-likelihood parameter estimate of the strength of a causal relationship, and allows us to define causal support for rate data. The first step in this analysis is defining a parameterization for Graph 0 and Graph 1 that is appropriate for outcomes that are rates rather than discrete trials.

The Poisson distribution is commonly used in statistics for modeling the number of events that occur within a fixed interval. Under this distribution, the number of events N occurring in a single unit of time will have probability

$$P(N) = e^{-\lambda} \frac{\lambda^N}{N!}, \tag{13}$$

where λ is the rate parameter. We can use the Poisson distribution to define a parameterization for Graph 0 and Graph 1 that is an extension of the linear and noisy-OR parameterizations to continuous time. Under a noisy-OR parameterization where the event E has parents B and C, the probability of E if just B is present is w_0, and the probability of E with both B and C present is $w_0 + w_1 - w_0 w_1$, where the last term corrects for double-counting the cases in which both B and C would have produced E. Extending this model to the case where events are emitted over a continuous interval, the probability of an event at any point in time is simply the sum of the probabilities for each of the parents, as in the linear parameterization, since the probability of two events from a Poisson process occurring simultaneously is zero. The resulting process is the sum of the Poisson processes associated with the parents, and the sum of two independent Poisson processes is a Poisson process with a rate equal to the sum of the rates of the original processes. This gives us the parameterization

$$P(N|b, c; \lambda_0, \lambda_1) = e^{-(b\lambda_0 + c\lambda_1)} \frac{(b\lambda_0 + c\lambda_1)^N}{N!}. \tag{14}$$

where λ_0 is the rate associated with B, and λ_1 is the rate associated with C.

Under the model specified by Eq. (14), ΔR is the maximum likelihood parameter estimate for λ_1. The correspondence to ΔP and causal power can be seen by taking the rate information as just the positive events in a contingency table where the total sample size is unknown, so $N(c^+) = NP(e^+|c^+)$ and $N(c^-) = NP(e^+|c^-)$ for unknown N. If we assume that N is fixed across different experiments, we can obtain estimates consistent with the ordering and magnitude implied by ΔP using $\Delta R = N(c^+) - N(c^-) = N\Delta P$. If we make the further assumption that N is very large, ΔR will also correspond to causal power, since $P(e^-|c^-)$ will tend to 1.

Using the parameterization given in Eq. (14), we can define causal support as in Eq. (10), where Graph 0 is the model in which B is the only parent of E, and Graph 1 is the model in which both B and C are parents of E. The details of this model are provided in Appendix A, where we also justify the approximation

$$\chi_r^2 = \frac{(N(c^+) - N(c^-))^2}{N(c^-)}. \tag{15}$$

This approximation bears the same relationship to causal support for rates as the Pearson χ^2 test for independence does for contingency data: it is a frequentist independence test that will be asymptotically equivalent to causal support. Computing χ_r^2 involves dividing the squared difference between the observed rates by the variance of the rate in the absence of the cause, comparing the magnitude of the effect of introducing the cause to the variation that should arise by chance. This may account for the efficacy of the grating contrast model used by Anderson and Sheu (1995), which involves a similar ratio.

Our analysis provides a set of models that generalize ΔP, causal power, causal support, and χ^2 to allow causal induction from rates. Unfortunately, the procedure used in previous experiments exploring causal induction from rates (Anderson & Sheu, 1995; Wasserman, 1990) prevents modeling of their data without detailed information about the performance of individual participants. These experiments used a procedure in which participants interacted with objects, and then observed whether there was an alteration in the rate of the effect after their interaction. The models described in this section cannot be applied to these data without a record of the number of periods of interaction and non-interaction. To address this issue we conducted our own experiments, in which participants were provided with information about the rate of occurrence of the effect in the presence and absence of the cause directly, using both summary (Experiment 4) and online (Experiment 5) formats.

13. Experiment 4

13.1. Method

13.1.1. Participants
Eighty-two Stanford University undergraduates took part in the study.

13.1.2. Stimuli
A questionnaire presented a summary of nine experiments involving different chemical compounds and electrical fields, giving the number of particle emissions inside and outside the electrical field. The number of particle emissions in each example was selected to give three critical sets of rates (expressed as $\{N(c^+), N(c^-)\}$ pairs): $\{52, 2\}$, $\{60, 10\}$, $\{100, 50\}$, for which $\Delta R = 50$, $\{12, 2\}$, $\{20, 10\}$, $\{60, 50\}$ for which $\Delta R = 10$, and $\{4, 2\}$, $\{12, 10\}$, $\{52, 50\}$, for which $\Delta R = 2$.

13.1.3. Procedure
The instructions outlined a hypothetical laboratory scenario:

> Imagine that you are working in a laboratory and you want to find out whether electrical fields influence the radioactive decay of certain chemical compounds. Below, you can see laboratory records for a number of studies. In each study, a

sample of some particular compound was placed inside a particular kind of electrical field for one minute, and the rate of radioactive decay was measured (in number of particles emitted per minute). Each study investigated the effects of a *different* kind of field on a *different* kind of chemical compound, so the results from different studies bear no relation to each other.

Of course, the chemical compounds can emit particles even when not in an electrical field, and they do so at different rates. Some compounds naturally decay at a fast rate, while others naturally decay at a slow rate. Thus, the decay rate of each compound was also measured for one minute in the absence of any electrical field. For each study below, you can see how many particles were emitted during one minute inside the electrical field, and during one minute outside of the electrical field. What you must decide is whether the electrical field increases the rate of particle emissions for each chemical compound.

Participants were instructed to provide ratings in response to a question like that of Experiment 2. Ratings were made on a scale from 0 (the field definitely does not cause the compound to decay) to 100 (the field definitely does cause the compound to decay). Each participant completed the survey as part of a booklet of unrelated experiments.

13.2. Results and discussion

The results are shown in Fig. 9, together with the model predictions. There was a statistically significant effect of $N(c^-)$ at $\Delta R = 50$ ($F(2,162) = 12.17$, $MSE = 257.27$, $p < .001$), $\Delta R = 10$ ($F(2,162) = 42.07$, $MSE = 468.50$, $p < .001$), and $\Delta R = 2$ ($F(2,162) = 29.87$, $MSE = 321.76$, $p < .001$). Causal support and χ_r^2 gave equivalent quantitative fits, $r = .978$, $\gamma = 0.35$, and $r = .980$, $\gamma = .01$, respectively, followed by grating contrast, $r = .924$, $\gamma = 0.43$, and ΔR, $r = .899$, $\gamma = 0.05$. Causal power assumes the wrong statistical model for this kind of data, but might be applied if we assumed that participants were comparing the rates to some hypothetical maximum number of particles that might be emitted, N. As N approaches infinity, causal power converges to ΔR. The trends predicted by causal power do not vary with the choice of N, so the value $N = 150$ was chosen to allow these trends to be illustrated. The predicted trends are clearly at odds with those observed in the data, reflected in the correlation $r = .845$, $\gamma = 0.06$.

Since ΔR predicts that responses within each of the critical sets should be constant, the statistically significant effect of $N(c^-)$ is inconsistent with this model. This trend is, however, predicted by causal support and χ_r^2. These predictions reflect the fact that the certainty in the value of λ_1 decreases as $N(c^-)$ increases: if the effect occurs at a high rate in the absence of the cause, it becomes more difficult to determine if an increase in the number of times the effect is observed when the cause is present actually reflects a causal relationship. The effect of $N(c^-)$ is thus a sign that people are attempting to determine the causal structure underlying their observations. To demonstrate that these results generalize to cases where rates are supplied perceptually rather than in summary format, we replicated this experiment using online presentation in Experiment 5.
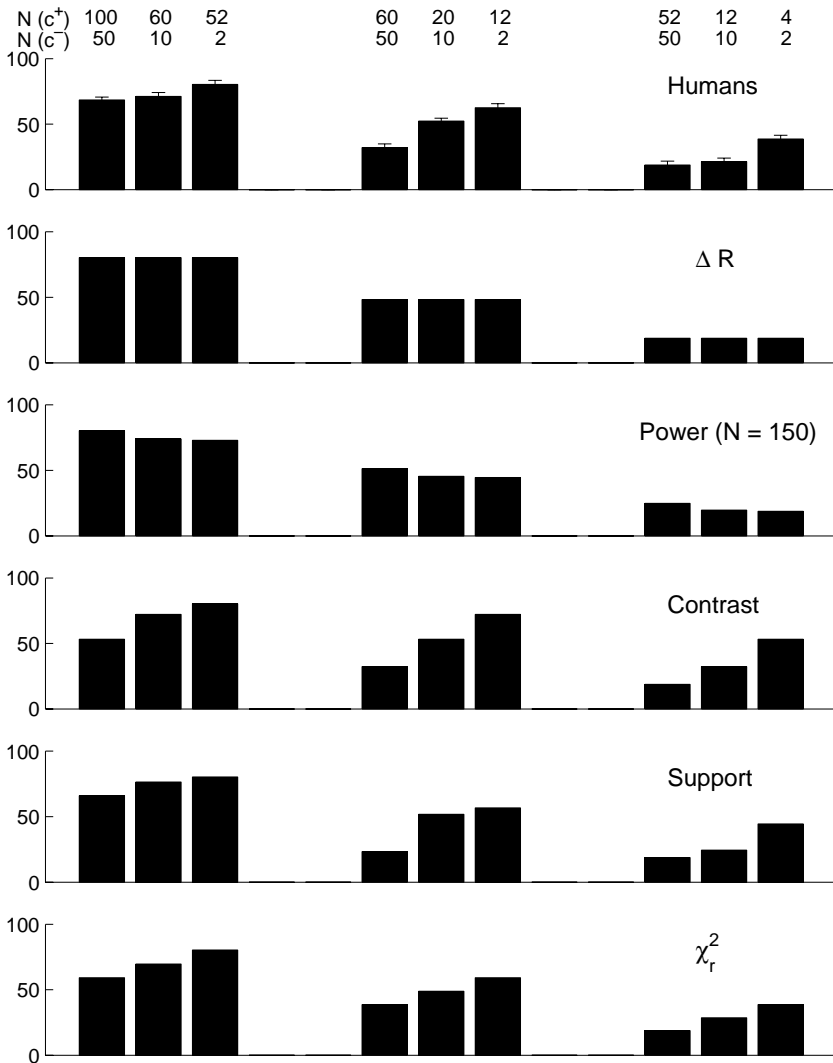
Fig. 9. Predictions of rational models compared with results of Experiment 4. Numbers along the top of the figure show stimulus rates, error bars indicate one standard error.

14. Experiment 5

14.1. Method

14.1.1. Participants

Participants were 40 members of the MIT Brain and Cognitive Sciences Department subject pool.

14.1.2. Stimuli

The stimuli were the same as those for Experiment 4, with the addition of three stimuli used for an initial practice phase. These stimuli used the $\{N(c^+), N(c^-)\}$ pairs $\{20, 28\}$, $\{10, 70\}$, and $\{16, 16\}$.

14.1.3. Procedure

The experiment was administered by computer. Participants were provided with instructions similar to those for Experiment 4, establishing the same laboratory cover story. The experiment consisted of three practice trials, followed by nine trials that used the same rate information as Experiment 3. In each trial, participants saw a picture of a novel mineral substance, and observed it for 30 s while it emitted particles, indicated by a beep and the temporary appearance of a mark on a "particle detector" shown on screen. They then clicked a button to turn on a magnetic field for another 30 s, and observed the particle emissions in the presence of the magnetic field. After having observed the mineral both in and out of the magnetic field, they rated the causal relationship between the magnetic field and particle emissions using a slider, with the same instructions as used in Experiment 4. The actual times of the particle emissions on each trial were generated from a uniform distribution over 30 s, with a minimum of 250 ms between emissions to ensure that they were perceptually distinct.

14.2. Results and discussion

The results are shown in Fig. 10, together with the model predictions. The results reproduced the trends found in Experiment 3, correlating with the previous data at $r = .975$. There was a statistically significant effect of $N(c^-)$ at $\Delta R = 50$ ($F_{(2,78)} = 5.82$, $MSE = 116.81$, $p < .005$) and $\Delta R = 10$ ($F_{(2,78)} = 23.18$, $MSE = 503.18$, $p < .001$), but not at $\Delta R = 2$ ($F_{(2,78)} = .97$, $MSE = 725.56$, $p = .385$). Causal support, χ_r^2, and ΔR all showed similar quantitative fit, $r = .951$, $\gamma = 0.416$, $r = .944$, $\gamma = 0.018$, and $r = .948$, $\gamma = 0.085$, respectively, followed by grating contrast, $r = .832$, $\gamma = 0.561$. Causal power, computed with $N = 150$, fared better on these data than on Experiment 4, $r = .912$, $\gamma = 0.047$.

The statistically significant differences among sets of rates for which ΔR is constant provides evidence against people's responses being driven by parameter estimation. The trends predicted by causal support are slightly less pronounced in these data than in Experiment 4, which we suspect may be a consequence of the more perceptual presentation format. In particular, the results for $\Delta R = 2$ may be a result of greater perceptual noise as $N(c^-)$ increases: for $\{4, 2\}$ and $\{12, 10\}$, it is quite easy to count the number of particle emissions in the presence and absence of the cause, and to make an inference informed by these counts. For $\{52, 50\}$, it is easy to lose count, and people may be less certain that the difference is small. Despite these small discrepancies, the results generally reproduce the trends observed in Experiment 4, and support the same conclusion: that people are making a structural inference when asked to assess a causal relationship, regardless of whether the data are discrete contingencies or dynamic rates. Our framework makes it possible to explain these results, and only causal support predicts human judgments for both of these kinds of data.
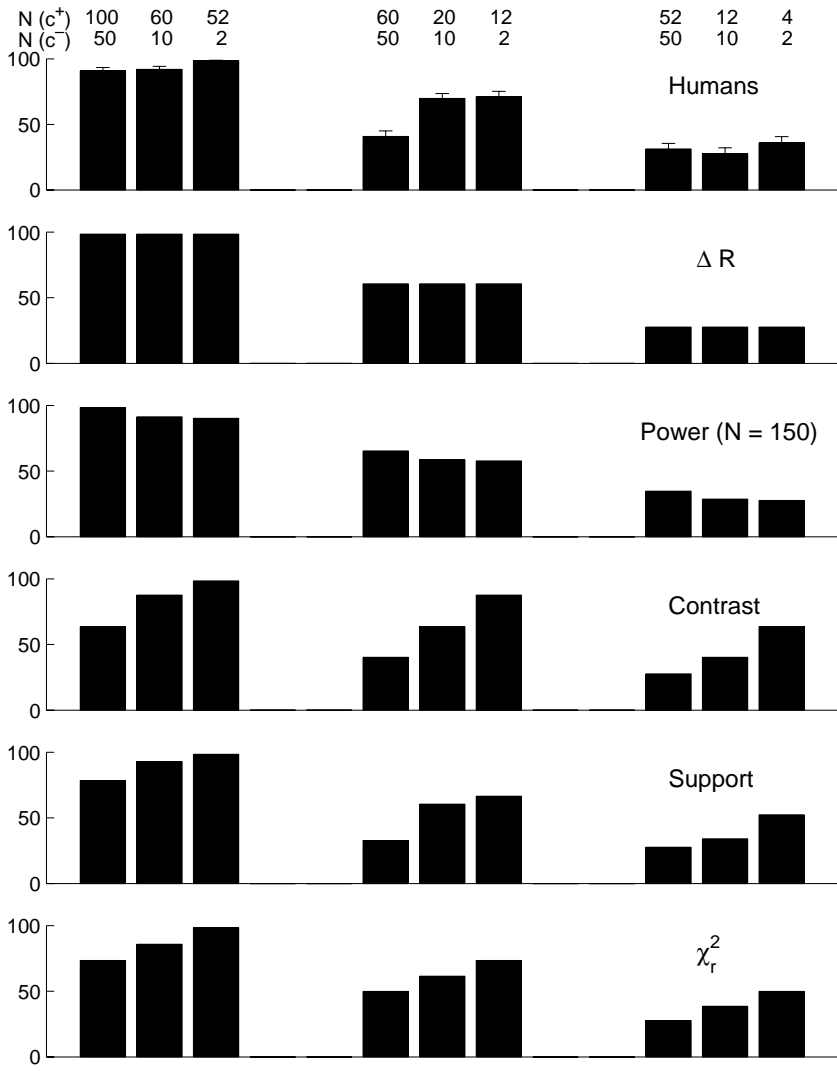
Fig. 10. Predictions of rational models compared with results of Experiment 5. Numbers along the top of the figure show stimulus rates, error bars indicate one standard error.

15. General discussion

We have presented a rational computational framework for analyzing the problem of elemental causal induction, using causal graphical models to emphasize the two components of this problem: parameter estimation—estimation of the strength of a causal relationship—and structure learning—evaluation of whether a causal relationship actually exists. Two leading rational models of elemental causal

induction, ΔP and causal power, both address the problem of parameter estimation. We have described five phenomena that suggest that, in many circumstances, people are approaching causal induction as structure learning. Causal support, a rational solution to the problem of learning causal structure, provides the best account of the two datasets at the center of the current debate about rational models—Buehner and Cheng (1997) and Lober and Shanks (2000)—and gives either the best or close to the best account of 10 out of the 12 other experiments we have analyzed (Anderson & Sheu, 1995; White, 1998, 2002c, 2003c). It also predicts several effects that no existing models can account for, in online, summary, and list formats, and in contingency and rate data, which we have validated through our own experiments.

Our results provide strong evidence that human judgments in causal induction tasks are often driven by the problem of structure learning, and that causal support provides an appropriate model of these judgments. We will close by discussing some issues raised by our analysis. First, we will describe some other models that address the problem of structure learning. We will then consider the circumstances under which we expect accounts based upon structure learning and parameter estimation to be most consistent with human judgments, before going on to clarify the relationship between causal support and ideas such as "reliability" (Buehner & Cheng, 1997; Perales & Shanks, 2003). Finally, we will sketch some ways in which our framework for elemental causal induction can be extended to shed light on other aspects of causal learning.

15.1. Structure learning in other models of causal induction

Our framework can be used to show that the leading models of elemental causal induction are both based upon parameter estimation. It also makes it possible to reinterpret several other psychological models in terms of structure learning. These models differ from causal support in terms of how they approach the issue of inferring causal structure, and in their assumptions about the functional form of the causal relationship between cause and effect. We will briefly discuss two models: Anderson's (1990; Anderson and Sheu, 1995) rational model, and White's (2002c, 2002b) proportion of Confirming Instances model.

15.1.1. Anderson's rational model

Anderson (1990) presented a Bayesian analysis of the problem of elemental causal induction. The model was quite complex, with seven free parameters, and a simpler model, with four free parameters, was presented by Anderson and Sheu (1995). The central idea of both models is that causal induction requires evaluating the hypothesis that the cause influences the effect. In our framework, this is a decision between Graph 0 and Graph 1, just as in causal support. Under Anderson's approach, these structures have a different parameterization from that used in causal support. Instead of the noisy-OR parameterization, a separate parameter is used to represent the probability of the effect for each set of values its causes take on. Either these parameters (Anderson & Sheu, 1995) or the prior on these parameters Anderson (1990) are chosen to provide the best fit to human data. As noted by Anderson

and Sheu (1995), this procedure results in predictions that are only slightly more restricted than linear regression using the cell counts from the contingency table (cf. Schustack & Sternberg, 1981). In contrast, causal support involves no free parameters, with the model predictions being computed directly from contingencies, as with ΔP and causal power.

15.1.2. The proportion of confirming instances

White (2000, 2002b, 2002a, 2002c, 2003a, 2003c) has also argued that people make judgments in causal induction tasks by assessing the evidence that a causal relationship exists. White (2002c) defined a measure of evidence termed the proportion of Confirming Instances (pCI), subsequently modified by White (2003a) to give the expression

$$\text{pCI} = \frac{N(e^+, c^+) + N(e^-, c^-) - N(e^+, c^-) - N(e^-, c^+)}{N(e^+, c^+) + N(e^-, c^-) + N(e^+, c^-) + N(e^-, c^+)}. \tag{16}$$

This equation also appears in other models of elemental causal induction (e.g., Catena et al., 1998), was originally proposed by Inhelder and Piaget (1958, p. 234), and is monotonically related to ΔP when $N(c^+) = N(c^-)$. The numerator of Eq. (16) has been explored as a model of causal induction in itself, being referred to as ΔD in the literature (e.g., Allan, 1980; Allan & Jenkins, 1983; Ward & Jenkins, 1965), and we show in Appendix A that it can be motivated from the perspective of structure learning. pCI produces the same predictions as ΔP for the majority of our experiments, and suffers many of the same problems. The exception is its sensitivity to sample size: as shown in Section 8, it does reasonably well in predicting the effect of varying $N(c^+)$ and $N(c^-)$, although it cannot account for the results of our Experiments 2 or 3. When learning from rates, the grating contrast measure defined by Anderson and Sheu (1995), given in Eq. (12), corresponds exactly to pCI.

15.2. Parameter estimation and structure learning in human judgments

Having distinguished between parameter estimation and structure learning as components of causal induction, it is natural to ask when we might expect one or the other to dominate human judgments. We have shown that across many experiments, causal support gives a better account of people's judgments than the maximum likelihood parameter estimate for the same model, causal power. These experiments all used tasks for which a structure-learning interpretation is reasonable. However, several recent studies have begun to use a different question format, asking people to give a counterfactual conditional probability, for which a strength estimate is appropriate (Buehner et al., 2003; Collins & Shanks, submitted). For example, rather than asking "Do injections cause gene expression?," we could ask "What is the probability that a mouse not expressing the gene before being injected will express it after being injected with the chemical?." Such questions elicit responses that are more consistent with causal power. These results can be explained using the framework we have established in this paper: causal power gives the maximum likelihood estimate of parameter w_1 in Eq. (4). The counterfactual questions used in

these experiments ask for the probability that, for a case in which C was not present and E did not occur, E would occur if C was introduced. Using the axiomatic treatment of counterfactuals developed by Pearl (2000), this probability is just w_1. Consequently, the appropriate way to answer counterfactual questions is via parameter estimation, and responses should correspond more closely to causal power than to causal support.

15.3. Causal support and reliability

Buehner and Cheng (1997) appealed to the notion of "reliability" in trying to explain aspects of their results that deviated from the predictions of the Power PC theory. To account for some of the data discussed above, they claimed that people sometimes conflate confidence in their estimates of causal power with the estimates themselves. Shanks and colleagues (Collins & Shanks, submitted; Perales & Shanks, 2003) have further explored this "reliability hypothesis." Our framework can be used to make the notion of reliability precise, and to explain why it might influence people's judgments.

Buehner and Cheng seem to consider reliability of secondary importance in evaluating causal relationships, acting something like the error bars on the estimate of causal power. Perales and Shanks (2003) convey a similar impression, equating reliability with confidence in assessments of the strength of a causal relationship. This notion of reliability thus seems to correspond to the certainty associated with a parameter estimate. Our framework provides a formal means of distinguishing between an estimate and its certainty, based upon the posterior distribution on w_1, as shown in Figs. 4 and 5. The location of the peak of this distribution indicates the strength of a relationship, and the width of this peak indicates the certainty in that estimate. Viewing causal induction as a structural inference makes it apparent that neither strength nor reliability should be considered primary: rational causal inferences should combine both factors. Causal support evaluates the evidence that w_1 differs from zero. This evidence generally increases as the peak of the posterior on w_1 moves away from zero (increasing strength), and as that peak becomes narrower (increasing reliability).

Our framework explains human judgments as a rational combination of strength and certainty, rather than the result of strength estimates being confounded by reliability. It also provides some suggestions about the circumstances in which the certainty in an estimate should be relevant. The results of Buehner et al. (2003) and Collins and Shanks (submitted) with counterfactual questions, summarized in Section 15.2, illustrate that certainty seems to influence judgments less when a task explicitly involves parameter estimation. Without the rational explanation provided by our framework, it is difficult to explain why judgments should be confounded less by reliability when people are asked counterfactual questions.

15.4. Limitations and extensions

The framework we have described in this paper, and the model of human judgments that we derived from it, causal support, address only the simplest cases of

causal induction. We have not discussed the dynamics of causal judgments in online studies, or more complex inductive tasks such as those involving multiple causes. We will briefly consider these limitations, and some potential directions for extending the framework.

15.4.1. The dynamics of causal judgments

The online presentation format makes it possible to ask participants to provide responses at several different points in the presentation of contingency information, providing the opportunity to measure the dynamics of causal judgments. Shanks et al. (1996) and López et al., 1998 discuss several phenomena involving changes in judgments over time, and trial order effects have been reported in several other studies (e.g.,Catena et al., 1998; Collins & Shanks, 2002). Specifying how structure learning and parameter estimation interact in people's judgments over time presents a particularly interesting problem for future research. We have conducted some preliminary work in this direction (Danks, Griffiths, & Tenenbaum, 2003), using a simple Bayesian formalism to provide a structure-sensitive strength estimate. This model also deals naturally with the distinction between generative and preventive causes, simultaneously updating a probability distribution over underlying models (generative, preventive, and no relationship) and distributions over the parameters of those models. This model provides an account of some of the qualitative features of learning curves from online experiments presented in Shanks et al. (1996).

15.4.2. Multiple causes

Many important phenomena in causal induction, such as backwards blocking Shanks (1985) and other retrospective revaluation effects (Shanks & Dickinson, 1987; Shanks et al., 1996) involve simultaneously learning about multiple causes. The framework for elemental causal induction that we have outlined in this paper can naturally be extended to accommodate multiple causes. The connection with causal graphical models clarifies how this extension should be made, as well as identifying some of the options for different modeling approaches. Steyvers et al. (2003) have shown that this kind of approach can be applied to situations involving multiple causes. We have also demonstrated that an approach based on causal graphical models can provide an explanation for backwards blocking phenomena in particular causal induction tasks (Tenenbaum & Griffiths, 2003).

15.4.3. Non-probabilistic causes

One of the consequences of distinguishing between parameter estimation and structure learning is a better understanding of causal induction with non-probabilistic causes. Approaches to causal induction based upon strength estimation only produce meaningful predictions when causes are probabilistic: if a causal relationship is deterministic, with the cause always producing the effect, the strength of the cause is 1.00. As a consequence, parameter estimation cannot be used to explain how people draw causal inferences about deterministic systems. Studies of causal induction from contingency data almost exclusively employ probabilistic causes, so this has not been identified as an issue for ΔP or causal power. However, in many studies in the

broader literature on causal induction people make graded inferences about non-probabilistic causes (e.g., Gopnik et al., 2004; Tenenbaum, Sobel, Griffiths, & Gopnik, submitted). Such inferences can be addressed within our framework: the decision between causal structures can be made regardless of causal strength, and the degree of evidence for a causal relationship can vary even with deterministic causes. Our approach to causal induction produces predictions consistent with people's judgments in several settings involving deterministic causes (e.g., Tenenbaum & Griffiths, 2003).

15.5. Conclusion

We have used causal graphical models to identify two components of the problem of causal induction: structure learning and parameter estimation. We have shown that previous rational models of causal induction only address the problem of parameter estimation. By emphasizing the role of structure learning in human causal induction, we have been able to explain a variety of phenomena that were problematic for previous models, and to understand the inference that underlies the discovery that a causal relationship actually exists. Our approach explains not only lay people's fundamental intuitions about cause and effect, but also the intuitions that drove discovery for early scientists, such as Dr. James Currie of the epigraph, and that continue to be important in the early stages of much contemporary scientific research.

Looking beyond the simple setting of elemental causal induction, our Bayesian approach provides a method for explaining how prior knowledge and statistical inference might be combined in causal learning. Human judgments about causality draw upon a rich body of knowledge about the mechanisms mediating between causes and effects, which relationships are plausible, and what functional forms particular relationships might take. The simple Bayesian inference that underlies our account of elemental causal induction can be augmented to capture the role of this knowledge, via constraints on the prior probabilities of particular causal relationships, and constraints on the functional form of the probability distributions defined on those structures. This capacity to capture the interaction between top-down knowledge and bottom-up statistical cues is one of the greatest strengths of our framework, and we are currently exploring how it might explain a broader range of causal inferences, including those that involve sparse data, hidden causes, and dynamical physical systems.

Appendix A

A.1. ΔP and causal power are maximum likelihood parameter estimates

Both ΔP and causal power are maximum likelihood estimates of the causal strength parameter for C in Graph 1 of Fig. 3(A), but under different parameterizations. For any parameterization of Graph 1, the log likelihood of the data is given by

$$\log P(D|w_0, w_1) = \sum_{e,c} N(e, c) \log P_1(e|c), \tag{17}$$

where D is contingency data $N(e, c)$, $P_1(e|c)$ is the probability distribution implied by the model, suppressing its dependence on w_0, w_1, and $\sum_{e,c}$ denotes a sum over all pairs of e^+, e^- and c^+, c^-. Eq. (17) is maximized whenever w_0 and w_1 can be chosen to make the model probabilities equal to the empirical probabilities:

$$P_1(e^+|c^+, b^+; w_0, w_1) = P(e^+|c^+),\tag{18}$$

$$P_1(e^+|c^-, b^+; w_0, w_1) = P(e^+|c^-).\tag{19}$$

To show that ΔP corresponds to a maximum likelihood estimate of w_1 under a linear parameterization of Graph 1, we identify w_1 in Eq. (6) with ΔP (Eq. (1)), and w_0 with $P(e^+|c^-)$. Eq. (6) then reduces to $P(e^+|c^+)$ for the case $c = c^+$ and to $P(e^+|c^-)$ for the case $c = c^-$, thus satisfying the sufficient conditions in Eqs. (18) and (19) for w_0 and w_1 to be maximum likelihood estimates. To show that causal power corresponds to a maximum likelihood estimate of w_1 under a noisy-OR parameterization, we follow the analogous procedure: identify w_1 in Eq. (4) with causal power (Eq. (2)), and w_0 with $P(e^+|c^-)$. Then Eq. (4) reduces to $P(e^+|c^+)$ for $c = c^+$ and to $P(e^+|c^-)$ for $c = c^-$, again satisfying the conditions for w_0 and w_1 to be maximum likelihood estimates.

A.2. ΔD, pCI, and structure learning

ΔD is approximately proportional to the log likelihood ratio in favor of Graph 1 if we define the parameterization of Graph 1 to be $P_1(e^+|c^+, b^+) = P_1(e^-|c^-, b^+) = \frac{1}{2} + \frac{1}{2}\epsilon$ for small ϵ, and the parameterization of Graph 0 to be $P_0(e^+|b^+) = \frac{1}{2}$. This choice of parameters gives

$$\frac{P_1(e^+|c^+, b^+)}{P_0(e^+|b^+)} = \frac{P_1(e^-|c^-, b^+)}{P_0(e^-|b^+)} = 1 + \epsilon,\tag{20}$$

$$\frac{P_1(e^-|c^+, b^+)}{P_0(e^-|b^+)} = \frac{P_1(e^+|c^-, b^+)}{P_0(e^+|b^+)} = 1 - \epsilon\tag{21}$$

and since $\log(1 + x) \approx x$, for values of x near 0, it follows that:

$$\log\frac{P(D|\text{Graph 1})}{P(D|\text{Graph 0})} \approx \epsilon[N(e^+, c^+) + N(e^-, c^-) - N(e^+, c^-) - N(e^-, c^+)].\tag{22}$$

However, the assumptions under which this unweighted combination of counts is an appropriate measure of the evidence for a causal relationship are overly restrictive: the background probability of the effect has to be 0.5, and the cause must be asymptotically weak. The definition of pCI, which normalizes this quantity by the sum of all entries in the contingency table, cannot be motivated from the perspective of structure learning, since it removes the effect of overall sample size.

A.3. Evaluating causal support

Causal support is defined as the log likelihood ratio in favor of Graph 1 over Graph 0:

$$\text{support} = \log \frac{P(D|\text{Graph 1})}{P(D|\text{Graph 0})}. \tag{23}$$

We obtain the likelihoods $P(D|\text{Graph 1}), P(D|\text{Graph 0})$ by integrating out the parameters w_0, w_1. This means that each value of the parameters is assigned a prior probability, and this probability is combined with the likelihood of the data given the structure and the parameters to give a joint distribution over data and parameters given the structure. We can then sum over all values that the parameters can take on, to result in the probability of the data given the structure. Thus, if we want to compute the probability of the observed data for the structure depicted by Graph 1, we have

$$P(D|\text{Graph 1}) = \int_0^1 \int_0^1 P_1(D|w_0, w_1, \text{Graph 1})P(w_0, w_1|\text{Graph 1})\mathrm{d}w_0\,\mathrm{d}w_1 \tag{24}$$

and the equivalent value for Graph 0 is given by

$$P(D|\text{Graph 0}) = \int_0^1 P_0(D|w_0, \text{Graph 0})\, P(w_0|\text{Graph 0})\, \mathrm{d}w_0, \tag{25}$$

where the likelihoods $P(D|w_0, w_1, \text{Graph 1})$, $P(D|w_0, \text{Graph 0})$ are specified by the parameterization of the graph, and the prior probabilities $P(w_0, w_1|\text{Graph 1})$, $P(w_0|\text{Graph 0})$ are set a priori. Integrating over all values of the parameters penalizes structures that require more parameters, simply because the increase in the dimensionality of the space over which the integrals are taken is usually disproportionate to the size of the region for which the likelihood is improved.

For generative causes, $P(D|\text{Graph 1})$ is computed using the noisy-OR parameterization, and for preventive causes, it is computed using the noisy-AND-NOT. We also need to define prior probabilities $P(w_0, w_1|\text{Graph 1})$ and $P(w_0|\text{Graph 0})$, to which we assign a uniform density. Because causal support depends on the full likelihood functions for both Graph 1 and Graph 0, we may expect causal support to be modulated by causal power, but only in interaction with other factors that determine how much of the posterior probability mass for w_1 in Graph 1 is bounded away from zero (where it is pinned in Graph 0). In the model for rate data, λ_0 and λ_1 are both positive real numbers, and priors for these parameters require different treatment. We take a joint prior distribution in which $P(\lambda_0) \propto \frac{1}{\lambda_0}$ is an uninformative prior, and $P(\lambda_1|\lambda_0)$ is Gamma$(1,\lambda_0)$.

A.4. An algorithm for computing causal support

Eq. (25) can be evaluated analytically. If w_0 denotes the probability of the effect occurring regardless of the presence or absence of the cause and we take a uniform prior on this quantity, we have

$$P(D|\text{Graph 0}) = \int_0^1 w_0^{N(e^+)}(1 - w_0)^{N(e^-)}\,\mathrm{d}w_0 = B(N(e^+) + 1, N(e^-) + 1), \tag{26}$$

where $B(r,s)$ is the beta function, and $N(e^+)$ is the marginal frequency of the effect. For integers r and s, $B(r,s)$ can be expressed as a function of factorials, being $\frac{(r-1)!(s-1)!}{(r+s-1)!}$. In general Eq. (24) cannot be evaluated analytically, but it can be approximated simply and efficiently by Monte Carlo simulation. Since we have uniform priors on w_0 and w_1, we can obtain a good approximation to $P(D|\text{Graph 1})$ by drawing m samples of w_0 and w_1 from a uniform distribution on [0,1] and computing

$$P(D|\text{Graph 1}) \approx \frac{1}{m} \sum_{i=1}^{m} P_1(D|w_{0i}, w_{1i}, \text{Graph 1}), \tag{27}$$

where w_{0i} and w_{1i} are the ith sampled values of w_0 and w_1. We thus need only compute the probability of the observed scores D under this model for each sample, which can be done efficiently using the counts from the contingency table. This probability can be written as

$$P_1(D|w_{0i}, w_{1i}, \text{Graph 1}) = \prod_{e,c} P_1(e|c, b^+; w_{0i}, w_{1i})^{N(e,c)}, \tag{28}$$

where the product ranges over e^+, e^- and c^+, c^-, and $P_1(e|c; w_{0i}, w_{1i})$ reflects the chosen parameterization—noisy-OR for generative causes, and noisy-AND-NOT for preventive. As with all Monte Carlo simulations, the accuracy of the results improves as m becomes large. For the examples presented in this paper, we used $m = 100{,}000$. Matlab code for evaluating causal support via Monte Carlo simulation can be obtained by writing to the corresponding author, or from the authors' webpages.

A.5. The χ^2 approximation

For large samples and weak causes we can approximate the value of causal support with the familiar χ^2 test for independence. There are both intuitive and formal reasons for the validity of the χ^2 approximation. Intuitively, the relationship holds because the χ^2 statistic is used to test for the existence of statistical dependency between two variables, and C and E are dependent in Graph 1 but not in Graph 0. A large value of χ^2 indicates that the null hypothesis of independence should be rejected, and that Graph 1 is favored. However, χ^2 assumes a different parameterization of Graph 1 from causal support, and the two will only be similar for large samples and weak causes.

The formal demonstration of the relationship between χ^2 and causal support is as follows. When the likelihood $P(D|w_0, w_1)$ is extremely peaked (e.g., in the limit $N \to \infty$), we may replace the integrals in Eq. (24) with supremums over w_0, w_1. That is, the marginal likelihood essentially becomes the maximum of the likelihood, and causal support reduces to the ratio of likelihood maxima—or equivalently, the difference in log likelihood maxima—for Graph 1 and Graph 0. Under these circumstances causal support reduces to the frequentist likelihood ratio statistic, equal to half of the G^2 statistic (e.g., Wickens, 1989). Correspondingly, Pearson's χ^2 for independence

$$\chi^2 = N \sum_{e,c} \frac{(P(e,c) - P(e)P(c))^2}{P(e)P(c)}, \tag{29}$$

can be shown to approximate twice causal support by a Taylor series argument: the second-order Taylor series of $\sum_i p_i \log \frac{p_i}{q_i}$, expanded around $p = q$, is $\frac{1}{2} \sum_i \frac{(p_i - q_i)^2}{q_i}$ (Cover & Thomas, 1991). The χ^2 approximation only holds when ΔP is small and N is large.

For learning with rates, the likelihood ratio statistic for comparing Graphs 0 and 1 under the parameterization given in Eq. (14) is

$$\chi_r^2 N(c^+) \log N(c^+) + N(c^-) \log N(c^-) - (N(c^+) + N(c^-)) \log \frac{N(c^+) + N(c^-)}{2}, \tag{30}$$

which, by essentially the same argument as that given above for G^2, will approximate the value of causal support in the large sample limit. Using the Taylor series argument employed in the contingency case, we obtain the χ^2 approximation given in Eq. (15), which holds only when the difference in rates is small.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15*, 147–149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin, 114*, 435–448.

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation, 14*, 381–405.

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

Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition, 23*, 510–524.

Buehner, M., & Cheng, P. W. (1997). Causal induction: The Power PC theory versus the Rescorla–Wagner theory. In M. Shafto & P. Langley (Eds.), *Proceedings of the19th annual conference of the cognitive science society* (pp. 55–61). Hillsdale, NJ: Lawrence Erlbaum Associates.

Buehner, M. J., 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*, 1119–1140.

Catena, A., Maldonado, A., & Cándido, A. (1998). The effect of the frequency of judgment and the type of trials on covariation learning. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 481–495.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory and Cognition, 18*, 537–545.

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

Cheng, P. W., & Holyoak, K. J. (1995). Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In H. L. Roitblat & J.-A. Meyer (Eds.), *Comparative approaches to cognitive science* (pp. 271–302). Cambridge, MA: MIT Press.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*, 545–567.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99*, 365–382.

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgment. *Memory and Cognition, 30*, 1138–1147.

Collins, D. J. & Shanks, D. R. (submitted). Conformity to the power PC theory of causal induction depends on type of probe question. *Memory and Cognition*.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 308–347.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Currie, J. (1798/1960). Medical reports on the effects of water, cold and warm, as a remedy in fever and other diseases, whether applied to the surface of the body or used internally, including an inquiry into the circumstances that render cold drink, or the cold bath dangerous in health. To which are added observations on the nature of fever and on the effect of opium, alcohol and inanition. In L. Clendening (Ed.), *Source book of medical history* (pp. 428–433). New York: Dover.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology, 47*, 109–121.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.). *Advances neural information processing systems* (Vol. 15, pp. 67–74). Cambridge, MA: MIT Press.

Danks, D. & McKenzie, C. R. M. (submitted). Learning complex causal structures.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In D. Fisher (Ed.), *Fourteenth international conference on machine learning* (pp. 125–133). San Francisco, CA: Morgan Kaufmann.

Friedman, N., & Koller, D. (2000). Being Bayesian about network structure. In *Proceedings of the 16th annual conference on uncertainty in AI, Stanford, CA* (pp. 201–210).

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines, 8*, 39–60.

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

Glymour, C., & Cooper, G. (1999). *Computation, causation, and discovery*. Cambridge, MA: MIT Press.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25*, 565–610.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 1–31.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 301–354). Cambridge, MA: MIT Press.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. London: Routledge and Kegan Paul.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, 79.

Jenner, E. (1798). *An inquiry into the causes and effects of the variolae vaccinae*.

Kass, R. E., & Rafferty, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Lagnado, D. A., & Sloman, S. (2002). Learning causal structure. In *Proceedings of the twenty-fourth annual meeting of the Cognitive Science Society*. London: Erlbaum.

Lober, K., & Shanks, D. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review, 107*, 195–212.

López, F. J., Cobos, P. L., Caño, A., & Shanks, D. R. (1998). The rational analysis of human causal and probability judgment. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 314–352). Oxford: Oxford University Press.

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

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner rule. comments on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1398–1410.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence, 56*, 71–113.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal inference. *Psychological Review, 111*, 455–485.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Pearson, K. (1904/1948). On the theory of contingency and its relation to association and normal correlation. In *Karl Pearson's early statistical papers* (pp. 443–475). Cambridge: Cambridge University Press.

Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgment. *Quarterly Journal of Experimental Psychology, 56A*, 977–1007.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141–1159.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Salmon, W. C. (1980). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110*, 101–120.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology, 97B*, 1–21.

Shanks, D. R. (1995a). Is human learning rational. *Quarterly Journal of Experimental Psychology, 48A*, 257–279.

Shanks, D. R. (1995b). *The psychology of associative learning*. Cambridge University Press.

Shanks, D. R. (2002). Tests of the power PC theory of cauasl induction with negative contingencies. *Experimental Psychology, 49*, 1–8.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 21, pp. 229–261). San Diego, CA: Academic Press.

Shanks, D. R., López, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilsitic contrast theories of human contingency judgment. *The psychology of learning and motivation* (Vol. 34, pp. 265–312). San Diego, CA: Academic Press.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.

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

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453–489.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.). *Advances neural information processing systems* (Vol. 13, pp. 59–65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In S. Becker, S. Thrun, & K. Obermayer (Eds.). *Advances in neural information processing systems* (Vol. 15, pp. 35–42). Cambridge, MA: MIT Press.

Tenenbaum, J. B., Sobel, D. M., Griffiths, T. L., & Gopnik, A. (submitted). Bayesian inference in causal learning from ambiguous data: Evidence from adults and children.

Vallee-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. *Quarterly Journal of Experimental Psychology, 51A*, 65–84.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the twentieth annual conference of the cognitive science society* (pp. 1102–1107). Mahwah, NJ: Erlbaum.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19*, 231–241.

Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the encironment. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 26, pp. 27–82). San Diego, CA: Academic Press.

Wasserman, E. A., Elek, S. M., Chatlosh, D. C., & Baker, A. G. (1993). Rating causal relations: Role of proabbility in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 174–188.

White, P. A. (1998). Causal judgement: Use of different types of contingency information as confirmatory and disconfirmatory. *European Journal of Cognitive Psychology, 10*, 131–170.

White, P. A. (2000). Causal judgment from contingency information: The interpretation of factors common to all instances. *Journal of Experimental Psychology: Learning Memory and Cognition, 26*, 1083–1102.

White, P. A. (2002a). Causal attribution from covariation information: the evidential evaluation model. *European Journal of Social Psychology, 32*, 667–684.

White, P. A. (2002b). Causal judgement from contingency information: Judging interactions between two causal candidates. *Quarterly Journal of Experimental Psychology, 55A*, 819–838.

White, P. A. (2002c). Perceiving a strong causal relation in a weak contingency: Further investigation of the evidential evaluation model of causal judgement. *Quarterly Journal of Experimental Psychology, 55A*, 97–114.

White, P. A. (2003a). Causal judgement as evaluation of evidence: The use of confirmatory and disconfirmatory information. *Quarterly Journal of Experimental Psychology, 56A*, 491–513.

White, P. A. (2003b). Effects of wording and stimulus format on the use of contingency information in causal judgment. *Memory and Cognition, 31*, 231–242.

White, P. A. (2003c). Making causal judgments from the proportion of confirming instances: the pCI rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 710–727.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Erlbaum.