Psychological Review
1997, Vol. 104, No. 2, 367–405

From Covariation to Causation: A Causal Power Theory

Patricia W. Cheng
University of California, Los Angeles

Because causal relations are neither observable nor deducible, they must be induced from observable events. The 2 dominant approaches to the psychology of causal induction—the covariation approach and the causal power approach—are each crippled by fundamental problems. This article proposes an integration of these approaches that overcomes these problems. The proposal is that reasoners innately treat the relation between covariation (a function defined in terms of observable events) and causal power (an unobservable entity) as that between scientists' law or model and their theory explaining the model. This solution is formalized in the power PC theory, a causal power theory of the probabilistic contrast model (P. W. Cheng & L. R. Novick, 1990). The article reviews diverse old and new empirical tests discriminating this theory from previous models, none of which is justified by a theory. The results uniquely support the power PC theory.

How does a reasoner come to know that one thing causes another? Psychological work on this issue of causal induction has been dominated by two basic approaches that have generally been regarded as opposing each other. One of these, the *covariation* approach, traces its roots to David Hume (1739/1987). This approach is motivated by the problem that the reasoners' sensory input—the ultimate source of all information that they have—does not explicitly contain causal relations. It follows that acquired causal relations must be computed from the sensory input in some way. One's sensory input clearly yields such information as the presence and absence of a candidate cause and of the effect, as well as the temporal and spatial relations between them. Treating such "observable" information as the input to the process of causal induction, models under this approach attempt in some way to assess covariation between a candidate cause and the effect (i.e., the extent to which the two vary together). An influential model of covariation—often called the contingency model—was proposed by researchers across various disciplines (Jenkins & Ward, 1965; Rescorla, 1968; Salmon, 1965). Interpreting this model in causal terms, ΔP_i, the *contingency* between candidate cause i and effect e is defined by

$$\Delta P_i = P(e|i) - P(e|\bar{i}), \qquad (1)$$

where $P(e|i)$ is the probability of e given the presence of i and $P(e|\bar{i})$ is that probability given the absence of i. (The conditional probabilities in the equation are estimated by the respective relative frequency of events for which e occurs in the presence and in the absence of i.) If ΔP_i is noticeably positive, i is a *generative* or facilitatory cause, and, if it is noticeably negative, i is an inhibitory or *preventive* cause. Otherwise, i is noncausal.

In the psychological literature, the Humean approach has split into several subdivisions. Statistical contingency models based on the ΔP rule have been contrasted with various types of associative models, in particular Rescorla and Wagner's (1972) discrepancy-based predictive learning rule (e.g., Anderson & Sheu, 1995; Baker, Berbrier, & Vallée-Tourangeau, 1989; Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Baker, Murphy, & Vallée-Tourangeau, 1996; Chapman, 1991; Chapman & Robbins, 1990; Cheng & Holyoak, 1995; Cheng & Novick, 1990, 1992; Dickinson, Shanks, & Evenden, 1984; Price & Yates, 1993; Shanks & Dickinson, 1987; Shanks, Lopez, Darby, & Dickinson, 1996; Wasserman, Elek, Chatlosh, & Baker, 1993; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996). They have also been contrasted with linear combination models (e.g., Arkes & Harkness, 1983; Downing, Sternberg, & Ross, 1985; Einhorn & Hogarth, 1986; Jenkins & Ward, 1965; Nisbett & Ross, 1980; Schustack & Sternberg, 1981; Shaklee & Tucker, 1980; Ward & Jenkins, 1965). It is important to emphasize, however, that all covariation models of causality face a major common hurdle: As many have noted, covariation does not always imply causation. ΔP is clearly insufficient as a criterion for causal induction, because not all covariational relations are perceived as causal. Many things follow one another regularly, yet one does not infer a causal relation between them. Sunrise might occur every day after a rooster on a farm crows (but sunrise does not occur at other times during the day when the rooster does not crow), and one class for a student might routinely follow another (but if the first class does not meet, for example during a holiday, neither does the second). Yet one would not infer that the rooster's crowing causes the sun to rise or that one class causes another. None of the subtypes of covariation models, as I show later, have provided an account

Preparation of this article was supported by National Science Foundation Grant DBS 9121298. I thank Keith Holyoak, Douglas Medin, Eric Melz, Mike Oaksford, David Shanks, and especially Clark Glymour and Laura Novick for their valuable comments on an earlier version of this article; and Woo-kyoung Ahn, Michael Waldmann, and Edward Wasserman for their valuable comments on a related article. I thank Ralph Miller and Thomas Wickens for highly informative discussion.

Parts of this article were presented at the September 1995 joint European Association of Experimental Social Psychologists/Society for Experimental Social Psychology meeting, Washington, DC.

Correspondence concerning this article should be addressed to Patricia W. Cheng, Department of Psychology, Franz Hall, University of California, Los Angeles, California 90095-1563. Electronic mail may be sent via Internet to cheng@lifesci.ucla.edu.

of why some perceived covariations are given a causal interpretation and others are not. Accordingly, a fundamental problem that remains for the covariation approach is: What takes one from covariation to causation?

This problem with the covariation approach has in part motivated the power approach, the alternative view that traces its roots to Kant (1781/1965). According to this approach, there exists some a priori knowledge that serves as a framework for interpreting input to the causal induction process. Kant proposed that people have the a priori knowledge that all events are caused: "All alterations take place in conformity with the law of the connection of cause and effect" (1781/1965, p. 218). Some psychologists have adopted Kant's proposal (Bullock, Gelman, & Baillargeon, 1982). More often, this view has been interpreted to mean that people do not infer that one thing is a cause of another unless they perceive or know of a specific generative source, causal mechanism, causal propensity, or causal power linking the candidate cause to the effect (e.g., Ahn & Bailenson, 1996; Ahn, Kalish, Medin, & Gelman, 1995; Bullock et al., 1982; Harré & Madden, 1975; Michotte, 1946/1963; Shultz, 1982; White, 1989, 1995).[1] *Causal power* (the general term I use to cover all of these variants) is the intuitive notion that one thing causes another by virtue of the power or energy that it exerts over the other (see Taylor, 1967, for a historical review and Cartwright, 1989, for a discussion of philosophical aspects of causal power). For example, when the sun warms one's back, one thinks of the sun as emitting energy, some of which reaches one's skin, raising its temperature. Likewise, when thoughtlessness is hurtful, one thinks of that thoughtlessness as having the power to produce pain. According to the power view, causes are not merely followed by their effects; rather, they produce or *generate* their effects. Sequences such as sunrise following crowing exhibit similar observable statistical characteristics as causal sequences but are missing the critical connection provided by the understanding of a causal power.

As just discussed, the problem of causal induction, which was first posed by Hume (1739/1987), has evolved into: How are causal relations constructed from the input that is available to one's information-processing system and distinguished from noncausal ones, including noncausal covariations (see Goodman, 1983)? Although the Kantian view has intuitive appeal, it does not address this problem. First, it suffers from the weakness of not being computational: It does not explicitly define a mapping between the ultimate input to the causal induction process and its output. Proponents of this view have not explained how the domain-independent knowledge that all events are caused can constrain inference from covariation in a specific domain. Second, with respect to the problem of causal induction, the interpretation of the Kantian view in terms of specific causal powers is crippled by its circularity. Specific causal powers are, by definition, causal. Moreover, they are nearly always acquired. In other words, this approach pushes the problem one step back but ultimately fails to solve it. The same problem arises with regard to the specific causal powers: Unless knowledge of these powers is innate, how do reasoners come to know them?

It is clear that neither approach is complete. I argue in this article, however, that each captures an element of truth—covariation is a component of the process of causal induction, and reasoners do have an a priori framework for interpreting input to that process. The theory I present integrates these approaches to overcome their individual problems.

Unexplained Empirical Phenomena of Causal Induction

In addition to the problems afflicting each of the two approaches to causal induction, another motivation for my theory is that there are several phenomena of natural causal induction that are inexplicable by any current psychological approach. Here I examine three such phenomena. These phenomena are not new or surprising. Rather, they are so mundane that they have generally been overlooked.

These phenomena all involve situations in which the effect occurs equally often in the presence of a candidate cause as in its absence (i.e., $\Delta P_i = P(e|i) - P(e|\bar{i}) = 0$), as do alternative causes (so that there is no confounding). First, consider evaluating whether a candidate cause produces an effect. For this goal, people cannot conclude that a *non-contingent* candidate cause (i.e., one for which $\Delta P = 0$) is noncausal when they know that an alternative cause is constantly producing the effect (so that $P(e|\bar{i}) = 1$). Let me illustrate this with an example. I suspected that I was allergic to certain foods, in reaction to which I had hives. When I went to the doctor, she made a grid of scratches on my back and put multiple samples of various foods on the scratched spots, one sample on each spot. After a few minutes, she observed that hives broke out at every spot (i.e., $P(e|i) = 1$, where e stands for hives and i stands for a food item). She might have concluded that I was allergic to every food tested. But it turned out that, in addition to being allergic to some food, I was also allergic to scratches. When my skin was scratched without being put in contact with any food, hives also broke out (i.e., $P(e|\bar{i}) = 1$). The tests for allergy to foods were therefore uninterpretable. Note that although ΔP equals 0 for every food in the test, the doctor did not conclude that I was not allergic to any of them.

This simple anecdote confounds all current psychological approaches to causal induction. Some covariational models predict that every candidate (every food item in this example) is noncausal (e.g., Cheng & Novick, 1990, 1992; Rescorla & Wagner, 1972) because the effect (hives) occurs just as often within a given context (scratching) when the candidate is present as when it is absent. Cheng and Holyoak's (1995) model explicitly assumes that reasoners would be uncertain under this situation but does not explain why. Other covariation models predict that the candidate is causal because the effect occurs very often in the presence of the candidate (e.g., the *a*-cell rule [see Nisbett & Ross, 1980, and Shaklee & Tucker, 1980], the *a*-minus-*b*-cell rule [see Shaklee & Tucker, 1980], and Schustack & Sternberg's model, 1981, given the parameter values reported in their article). The power view does not explain why scratching should block an inference regarding foods. In sum, no model explains what seems to be a reasonable answer, that the reasoner cannot reach a conclusion.

[1] Kant (1781/1965), in fact, argued that people have general, but not specific, a priori knowledge about causality. He wrote, "certainly, empirical laws, as such, can never derive their origin from pure understanding" (p. 148) and to "obtain any knowledge whatsoever of these special laws, we must resort to experience" (p. 173).

A second phenomenon involves the difference between causal judgments in two situations, in both of which $\Delta P = 0$. For such situations, a distinction is required between (a) the case in which the effect occurs only sometimes or never occurs within a certain context (i.e., $P(e|i) = P(e|\bar{i}) < 1$), and (b) the case illustrated by the earlier anecdote in which the effect always (or nearly always) occurs within a certain context (i.e., $P(e|i) = P(e|\bar{i}) \cong 1$). In the first situation, the candidate is judged to be noncausal at asymptote (e.g., Baker et al., 1989; Cheng & Novick, 1990; Fratianne & Cheng, 1995; Shanks & Lopez, cited in Shanks, 1995; Waldmann & Holyoak, 1992). In the second situation, however, the candidate is judged to have an uncertain causal status when participants are given the option of explicitly expressing uncertainty (Fratianne & Cheng, 1995). When participants are not given this option, the candidate receives a rating about midway between being a cause and being noncausal (Chapman, 1991; Chapman & Robbins, 1990; Shanks, 1985b, 1991; Shanks & Dickinson, 1987; Waldmann & Holyoak, 1992; Williams, Sagness, & McPhee, 1994). Providing further evidence for this difference, Fratianne and Cheng (1995, Experiment 3) and Waldmann and Holyoak (1992, Experiment 3) found the difference in causal judgments under these two situations to be highly reliable when they directly compared them in a single experiment. The difference in causal judgments in the two situations has never been explained.

A third unexplained empirical phenomenon involves the evaluation of the inhibitory nature of candidate causes. Such evaluations have a constraint that is diametrically opposite to that for the evaluation of the generative nature of a candidate: Reasoners cannot conclude that a noncontingent candidate cause is not inhibitory when they know that no generative cause is present in the context, and hence the effect never occurs in the first place (i.e., $P(e|i) = P(e|\bar{i}) \cong 0$). Consider this hypothetical example. Suppose that you are a medical researcher who has developed a drug for relieving headaches. You ask an assistant to conduct an extensive test of this drug. He or she randomly assigns participants to two groups, administering the drug to one group (the experimental group) and a placebo to the other (the control group). After observing the participants, your assistant informs you that there is no difference in the occurrence of headaches between the experimental and control groups after the respective treatments. He or she concludes that the drug is ineffective. You examine the results closely and find, to your surprise, that participants in the experimental group did not get headaches after receiving the drug. You see that your assistant was right, however, that there was no difference between the two groups—participants in the control group likewise did not get headaches, either before or after receiving the placebo! Would you agree with your assistant that the drug is ineffective for relieving headaches? Your answer is likely to be that your assistant conducted an absurd test, the results of which are incapable of being informative: Testing the ability of a drug to relieve (i.e., inhibit) some symptom requires that at least some participants exhibit that symptom before the administration of the drug. Remarkably, no current model of causal induction can explain the reasonable conclusion that the study is uninformative.

Boundary Conditions on Covariation Models

As has been shown, even when there is no confounding by alternative causes, different causal judgments may result from a contingency of zero. A noncontingent candidate will be interpreted differently depending on the goal of the inference and the context in which the zero contingency occurs. I have considered the following situations: (a) when an alternative cause is known to be present and is always producing the effect (so that $P(e|i) = P(e|\bar{i}) \cong 1$); (b) when an alternative cause is present but is producing the effect only sometimes (so that $0 < P(e|i) = P(e|\bar{i}) < 1$); and (c) when no alternative cause is present and the effect never occurs (so that $P(e|i) = P(e|\bar{i}) \cong 0$). If one is evaluating whether a candidate cause produces an effect, it is not possible to draw a firm conclusion about a noncontingent candidate when $P(e|\bar{i}) \cong 1$ (i.e., in the first situation), but one would infer that such a candidate does not produce the effect for other values of $P(e|\bar{i})$. If one is evaluating whether a candidate cause prevents an effect, however, the conclusions regarding the first and third situations are exactly reversed: One would conclude that a noncontingent candidate is not an inhibitory cause when $P(e|\bar{i}) \cong 1$, whereas no firm conclusion can be drawn about such a candidate when $P(e|\bar{i}) \cong 0$. In the second situation, when the value of $P(e|\bar{i})$ is between these extremes, one would conclude that a noncontingent candidate is neither a generative nor an inhibitory cause. (The conditioning literature [Baker et al., 1996; Miller, Barnet, & Grahame, 1995; Williams, 1996] shows parallels of these radical asymmetries between generative and inhibitory causes at the two extreme values of $P(e|\bar{i})$.)

These systematic variations in inferences for noncontingent candidates suggest boundary conditions for covariation models. Why are there boundary conditions? And why is a boundary condition for assessing generative causal power diametrically opposite to one for assessing preventive power?

A Resolution Between the Two Views: Causal Power Is to Covariation as Theory Is to Model

Laws and models in science, which deal with observable properties, are often explained by theories, which posit unobservable entities. In chemistry, for example, the kinetic theory of gases explains gas laws such as Boyle's law (pressure · volume = constant) and their boundary conditions (e.g., when temperature and the number of moles of gas are held constant for Boyle's law) by positing gases as tiny particles in a large space moving at a speed proportional to their temperature. The bombardment of the particles on the container walls yields the gas laws.

I propose that causal power is to covariation as the kinetic theory of gases is to Boyle's law. When ordinary folks induce the causes of events, they innately act like scientists in that they postulate unobservable theoretical entities (in this case, an intuitive notion of causal power) that they use to interpret and explain their observable models (in this case, their intuitive covariation model). That is, people do not simply treat observed covariations as equivalent to causal relations; rather, they interpret their observations of covariations as manifestations of the operation of unobservable causal powers, with the tacit goal of estimating the magnitude of these powers. The idea that people are intuitive scientists is by no means new (Kelley, 1967, 1973; Nisbett & Ross, 1980), but I extend this analogy to refer to the explaining of a model by a theory. I formalize my proposal by postulating a mathematical concept of causal power and deriving a theory of a model of covariation based on that concept. I do not claim that people either understand or use the mathematical version of my theory, only that they implicitly use a qualitative version of it (Cheng & Fratianne, 1995).

Scope

In this article, I assume that the topic of causal inference may be divided into two component issues: how an acquired causal relation is first induced and how prior domain-specific causal knowledge (whether innate or learned) regarding superordinate kinds influences subsequent causal judgments (Cheng, 1993; Kelley, 1967; Morris & Larrick, 1995). A child might learn from experience that dropping a ceramic plate on the ground breaks it, whereas dropping a plastic plate does not, and that touching a soap bubble causes it to pop, whereas touching a balloon does not. These examples illustrate the process of induction. The sunrise example given earlier illustrates the influence of prior domain-specific superordinate knowledge. The intuition is that one cannot conceive of any causal power or mechanism underlying crowing that produces sunrise. For example, based on experience with sounds that one can manipulate, a reasoner might have induced that sounds in general do not elevate objects: Tables do not elevate in the midst of heated arguments, and people do not elevate in the midst of a fire alarm. A rooster's crowing does not elevate even a straw in the barn; it therefore surely does not elevate the sun. Sound, therefore, is not a viable causal power with respect to sunrise. Neither is any other candidate that crowing calls to mind.

This article addresses only the issue of causal induction. The assumption that causal induction and the influence of domain-specific prior causal knowledge are separable processes is justified by numerous experiments in which the influence of such knowledge can be largely ignored (e.g., Baker et al., 1993; Cheng & Novick, 1990; Shanks, 1991; Wasserman, Chatlosh, & Neunaber, 1983; Wasserman et al., 1993). The results of these experiments demonstrate that the induction component can indeed operate independently of prior causal knowledge. The assumption of separability does not imply that these processes cannot jointly influence a particular causal judgment. One's visual system, for example, clearly has a separable component that processes external input, but the separability of this component does not imply that external input and prior knowledge cannot (or do not) jointly influence a perceptual episode.

Even within the scope of causal induction, this article does not address an often-raised question: How are effects and their candidate causes selected out of the indefinitely large pool of possible representations? This is a fundamental issue that requires extensive further study.

In sum, this article focuses on how people induce causal relations without the benefit of domain-specific causal knowledge when candidate causes and effects are clearly defined. To my knowledge, no solution that is free of the problems noted earlier has yet been offered.

Overview

In the rest of this article, I first review the probabilistic contrast model (Cheng & Holyoak, 1995; Cheng & Novick, 1990, 1991, 1992; Melz, Cheng, Holyoak, & Waldmann, 1993) and present a mathematical formulation of a power theory of this model. To distinguish this power theory from previous interpretations of the power view, I refer to it as the *power PC theory.*[2]

Second, I analyze Rescorla and Wagner's (1972) model (R–W model from here on), the dominant alternative model to which the probabilistic contrast model and the traditional contingency model have been compared. This model is the most influential associationist model of Pavlovian conditioning of the past quarter century. Because findings about conditioning appear to have close parallels in human inference (e.g., Dickinson et al., 1984; Rescorla, 1988), this model has recently been adopted more generally as a model of learning, categorization, and causal inference (e.g., Baker et al., 1993, 1996; Chapman & Robbins, 1990; Gluck & Bower, 1988; Price & Yates, 1993; Shanks, 1995; Shanks & Dickinson, 1987; Wasserman et al., 1993, 1996; see Siegel & Allan, 1996, for a review of the influence of this model). In addition to this model's prominence in Pavlovian conditioning and causal induction, another reason I am focusing on the model is that the learning rule it incorporates is a version of the "delta rule" commonly used in connectionist models (e.g., Kruschke, 1992, 1993; McClelland & Rumelhart, 1985; Rumelhart & McClelland, 1986). My analysis of this model should therefore be relevant to connectionist models using this rule, whatever the content domain of the model.

The central difference between my approach and the associationist approach involves the distinction between the computational and algorithmic levels of cognitive analysis (Marr, 1982). Marr argues that the issues of what function is being computed by an information process and why it is computed (i.e., the goal and the constraints that motivate the process) logically precede the issue of how a given function is computed. He classifies the former issues as being at a computational level and the latter as being at an algorithmic level. If the function being computed does not mirror its human counterpart, no algorithm that computes this function can possibly be an accurate model of human cognition.

The R–W model was founded on an algorithm for discrepancy reduction on a trial-by-trial basis. Although an algorithmic model does compute some function at the computational level, such functions are almost never specified. One consequence is that their properties at the computational level are rarely systematically examined. In contrast, the power PC theory is a computational-level explanation; it seeks to specify the abstract function relating the input and the output of the process of causal induction given the constraints that govern the problem of causal induction. In this article, I provide an analysis of what function the R–W algorithm asymptotically computes so that it can be compared with the power PC theory at Marr's (1982) computational level. Because the R–W model is equivalent to the least-mean-squares rule of Widrow and Hoff (1960; see Sutton & Barto, 1981), one might think that it is normative. My analysis of this model in terms of causal power shows when and why it is not.

Third, I evaluate the power PC theory and the R–W model

[2] Neil Cheng Holyoak suggested this nickname for my theory when I complained that "the causal power theory of the probabilistic contrast model" is a mouthful. "Power PC" is also the name of Neil's favorite personal computer.

against old and new empirical results regarding causal induction. These include the basic influence of contingency (e.g., Allan & Jenkins, 1980, 1983; Baker et al., 1989; Shanks, 1985a, 1987; Wasserman et al., 1983, 1993); the subtle but systematic influence of the base rate of the effect (i.e., $P(e|\bar{i})$) on the magnitude of causal judgments for candidates with a given ΔP_i (Allan & Jenkins, 1983; Wasserman et al., 1983, 1993); the distinction between causes and enabling conditions (Cheng & Novick, 1991); the distinction between a novel candidate and an irrelevant one (e.g., Baker & Mackintosh, 1976), the causal counterpart of blocking and induced overshadowing in Pavlovian conditioning (e.g., Baker et al., 1993; Chapman & Robbins, 1990; Dickinson & Burke, 1996; Fratianne & Cheng, 1995; Price & Yates, 1993; Shanks, 1991; Waldmann & Holyoak, 1992), of overexpectation and the absence of it (Park & Cheng, 1995), and of conditioned inhibition and the direct and indirect extinction of such inhibition (Williams, 1995, 1996; Yarlas, Cheng, & Holyoak, 1995); the asymmetry in the interpretation of zero contingencies between the induction of generative and preventive causes manifested in blocking, conditioned inhibition and its extinction, overexpectation, and the influence of $P(e|\bar{i})$ on the magnitude of causal judgments for candidates with the same ΔP_i; the retrospective nature of some causal judgments (e.g., Chapman, 1991; Williams et al., 1994; Yarlas et al., 1995); preasymptotic performance (e.g., Dickinson & Burke, 1996; Shanks, 1985a); the influence of the utility of the outcome (Chatlosh, Neunaber, & Wasserman, 1985, Experiment 2); and trial-order effects and other findings involving ambiguous causal estimates (e.g., Chapman, 1991; Shanks, 1991, Experiment 3; Shanks, 1995; Wagner, Logan, Haberlandt, & Price, 1968; Williams et al., 1994).

Fourth, I review linear combination models, showing that they are clearly inaccurate as models of causal induction. Despite their many shortcomings, such models aptly describe a robust finding that indisputably contradicts both the probabilistic contrast model and the asymptotic predictions of the R–W model: Reasoners tend to weight frequencies of the effect in the presence of a candidate cause more than those in its absence (e.g., Anderson & Sheu, 1995; Baron, 1994; Dickinson & Shanks, 1986; Schustack & Sternberg, 1981; Wasserman et al., 1993). I show that this apparent bias follows from the normative power PC theory.

My analysis and review lead me to conclude that (a) the probabilistic contrast model (Cheng & Novick, 1992) gives the best description of the model of natural causal induction in the reasoner's head but that this model in the head has boundary conditions, (b) these boundary conditions are explained by the reasoner's theory of causal power (some situations allow estimation of causal power, whereas others inherently do not), and (c) reasoners interpret the mapping between the observable input (e.g., the presence and absence of a candidate cause and of the effect) and the explicit output of their model (the value of ΔP) according to their theory of causal power. Finally, I show how the power PC theory overcomes some of the fundamental problems that cripple the two approaches to causal induction.

The Probabilistic Contrast Model

It has long been argued that covariation is a component of the normative criterion for inferring a causal link between a factor and an effect (e.g., Kelley, 1967; Salmon, 1965). This criterion, however, has been criticized for not being descriptive of natural causal induction. Deviations from various putatively normative models of covariations have been reported in social psychology, cognitive psychology, and philosophy. For example, philosophers (Hart & Honoré, 1959/1985; Mackie, 1974; Mill, 1843/1973) have noted that when one considers all events related to an effect (e.g., a forest fire), one finds that the effect is almost invariably produced by multiple factors that are individually necessary and jointly sufficient to produce the effect (e.g., the dropping of a lit cigarette, the presence of oxygen, or the combustibility of the trees). Despite the apparently equal logical status of the contributing factors, within a certain context people might infer that a single covarying factor (e.g., the cigarette) is the cause, whereas other factors (e.g., oxygen and trees) are merely enabling conditions. Moreover, what is perceived to be an enabling condition in one context can become a cause in a different context. For example, Hart and Honoré (1959/1985, p. 35) wrote that "if a fire breaks out in a laboratory or in a factory, where special precautions are taken to exclude oxygen during part of an experiment or manufacturing process . . . there would be no absurdity at all in *such* a case in saying that the presence of oxygen was the cause of the fire." Likewise, the causal attribution literature in social psychology has reported that people suffer from a variety of biases (for reviews, see Cheng & Novick, 1990; Jaspars, Hewstone, & Fincham, 1983; Nisbett & Ross, 1980). For example, people have a bias toward attributing effects to a person rather than to other factors (e.g., a situation) that apparently have equal objective status.

Cheng and Novick (1990) proposed the probabilistic contrast model as a generalized contingency model to provide a descriptive account of the use of statistical regularity in natural causal induction. The model, which applies to events that can be represented by discrete variables, assumes that an initial criterion for identifying potential causes is perceived priority (causes must be understood to precede their effects). A potential cause is then evaluated by its contingency computed over a *focal set*, which is a contextually determined set of events that the reasoner uses as input to the covariation process. It is often not the universal set of events, contrary to what had been assumed by philosophers.

Factors sometimes combine in a nonindependent way to produce the effect, as when the dropping of a lit cigarette and the dryness of the forest jointly produce a forest fire or when talent and hard work jointly produce success. Such situations involve conjunctive causes. According to the probabilistic contrast model, such causes are evaluated via interaction contrasts. Here, I do not review this model's explanation of conjunctive causes but, instead, focus on simple causes that produce the effect independently of other causes. A simple candidate cause that consists of a single factor is evaluated by a main-effect contrast within a current focal set; using the events in that set, such a contrast for evaluating a candidate cause i of effect e is defined as in Equation 1. (I use the terms *contrast* and *contingency* interchangeably.)

Confidence in the assessment of a contrast is presumed to increase monotonically with the number of cases observed (Cheng & Holyoak, 1995). In addition to influencing confidence, number of observations changes the denominators of the relative frequencies used to estimate the two conditional probabilities. The latter influence explains the differentiation between a novel candidate cause and an irrelevant one, a phenomenon that has a parallel in the classical conditioning literature: learned irrelevance (Baker & Mackintosh, 1976, 1977; Kremer, 1971; see Cheng & Holyoak's, 1995, explanation).

The concept of a focal set explains the distinction people make between a cause and an enabling condition. In addition to defining a cause, Cheng and Novick (1991, 1992) defined an enabling condition: Candidate *i* is an enabling condition for a cause *j* if *i* is constantly present in a reasoner's current focal set but covaries with the effect *e* in another focal set, and *j* no longer covaries with *e* in a focal set in which *i* is constantly absent. As mentioned, the focal sets used by ordinary folk are often not the universal set. To illustrate with the forest fire example, because fire occurs more frequently given the dropping of a lit cigarette than otherwise, the dropped cigarette is a cause. Oxygen, however, is present in all forests. Its contrast therefore cannot be computed within the current focal set. It is not causally irrelevant, however, because it does covary with fire in another focal set, one that includes events in which oxygen is absent as well as those in which it is present (e.g., in chemistry laboratories). Oxygen is therefore an enabling condition. Finally, it is an enabling condition rather than an alternative cause because, in yet another focal set in which oxygen is always absent (e.g., also in chemistry laboratories), a lit cigarette no longer covaries with a bigger fire.

Similarly, the probabilistic contrast model explains the inductive biases reported in the social psychology literature by a discrepancy between the focal sets assumed by the investigators and the participants. The focal set for the investigators—the "universal" set consisting of all of the events presented in an experiment—is often too narrow for the participants, because participants often recruit relevant events from their prior experience for inclusion. To test this hypothesis, my collaborators and I manipulated the participants' focal sets in two ways: (a) We specified and varied information that had previously been assumed to be irrelevant by the investigators (Cheng & Novick, 1990), and (b) we left such information unspecified but manipulated and measured participants' assumptions about it (Novick, Fratianne, & Cheng, 1992). We found that our manipulations produced variations in causal judgment, variations that have been regarded as biases but are, in fact, systematically predicted by the computation of contrast over a more accurate focal set.

In addition to explaining many biases in social causal inference and the distinctions among a single-factor cause, a conjunctive cause, an enabling condition, a novel candidate cause, and an irrelevant factor, the probabilistic contrast model also explains a set of phenomena sometimes termed cue competition: the influence of alternative causes on the evaluation of a candidate cause. To explain cue competition, Cheng and Holyoak (1995; Melz et al., 1993) added auxiliary assumptions to Cheng and Novick's (1992) model. These assumptions specify the focal sets reasoners prefer for computing contrast: those in which plausible alternative causes are controlled (cf. Cartwright, 1989; Salmon, 1980). Preferences for how they are controlled, however, differ depending on the nature of the assessment: To assess whether a candidate cause is generative, reasoners prefer to compute contrast conditional on the absence of all other plausible causes; to assess whether a candidate is preventive, however, they prefer to compute contrast conditional on some generative cause being constantly present. In addition to specifying the optimal conditional contrasts, these auxiliary assumptions specify when a contrast is uninterpretable, even while alternative causes are controlled. For example, one of these assumptions corresponds to the concept of a "ceiling effect" in experimental design. This list of assumptions has sometimes been regarded as complex or unprincipled (e.g., Baker et al., 1996; Shanks, 1993).

Cheng and Holyoak's (1995) model, which is adapted from proposals in philosophy and principles of experimental design, is an attempt at specifying when covariation implies causation. Inherited from its parents in philosophy (Cartwright, 1989; Salmon, 1980), however, is a serious shortcoming: Because inference regarding a candidate cause is based on conditional contrasts, it is dependent on knowledge about alternative causes. This leads to a problem of how one infers the alternative causes. To overcome this problem of how inference begins, Cheng and Holyoak assumed an initial associationist stage. This associationist heuristic has the unfortunate side effect of undermining the justification for inferring causation from the obtained covariation.

Explaining the Probabilistic Contrast Model by a Theory of Causal Power

In this section, I present an explanation of the probabilistic contrast model (Cheng & Novick, 1990, 1992) in terms of causal power. My explanation shows (a) the conditions under which this model provides an estimate of causal power and (b) how well it does so under those conditions. It assumes that the reasoner believes that *there are such things in the world as causes that have the power to produce an effect and causes that have the power to prevent an effect and that only such things influence the occurrence of an effect* (cf. Bullock et al., 1982; Kant, 1781/1965). I first explain the probabilistic contrast model in terms of the theoretical concept of generative causal power. I then explain the same model in terms of preventive causal power. As I derive my results, I interpret them from the point of view of a reasoner who infers the magnitude of the unobservable causal power from observable events based on his or her theoretical explanation. I also consider special cases in which he or she holds beliefs about alternative causes.

Main-Effect Contrast and Generative Causal Power

To evaluate whether a candidate cause produces an effect, my theory explains non-negative main-effect contrasts in terms of generative causal power.[3] My analysis assumes that the power of a cause x to produce an effect e can be represented by p_x, the probability with which x produces e when x is present. Thus, $0 \leq p_x \leq 1$ for all x. Whereas $P(e|x)$—the probability of e occurring in the presence of x—can be directly estimated by observable events (the relative frequency of e occurring in the presence of x), p_x—the power of x—is a theoretical entity that

[3] Clark Glymour clarified my theory.

can only be indirectly estimated. (All conditional probabilities are represented by uppercase *P*s because they can be estimated directly by observable events; all causal powers are represented by small letter *p*s because they are theoretical entities and are not directly observable.) $P(e \mid x)$ coincides with p_x when no other cause is present or exists. They are not, however, equal in general. This is because other causes, known or unknown to the reasoner, might be present when x is present. For example, the probability with which birth defects occur in the offspring of women who drink alcohol during pregnancy ($P(\mathit{defects} \mid \mathit{alcohol})$) is not in general equal to the probability with which alcohol produces birth defects ($p_{alcohol}$), because other causes of birth defects might be present. How, then, does one estimate the causal power of a candidate cause in question? To do so, I distinguish between i, the candidate cause, and the composite of (known and unknown) causes alternative to i, which I designate as a.[4] $P(e \mid i)$ and p_i are not equal if a is present and produces e in the presence of i (i.e., if $P(a \mid i) \cdot p_a \neq 0$).

Let me summarize my assumptions in terms of my notations:

1. when i occurs, it produces e with probability p_i; when a occurs, it produces e with probability p_a; and nothing else influences the occurrence of e;
2. i and a influence the occurrence of e independently; and
3. i and a influence the occurrence of e with causal powers that are independent of how often i and a occur (e.g., the probability of e being produced by a, that is, the probability of the intersection of a occurring and a producing e, is $P(a) \cdot p_a$].

To evaluate whether i produces e, I first show how the theoretical entities p_i and p_a explain an observable ΔP_i. According to my theory, the reasoner theorizes that e can be produced by i or a (independently)—that is, the event "e occurring" is the union of two independent events: e produced by i and e produced by a. It follows from probability theory that $P(e)$, the probability of the union, is the sum of the probabilities of the constituent events minus their intersection:

$$P(e) = P(i) \cdot p_i + P(a) \cdot p_a - P(i) \cdot p_i \cdot P(a) \cdot p_a. \quad (2)$$

To explain ΔP_i in terms of this theory, consider separately the probability of e occurring when i is present and when i is absent. First, to explain $P(e \mid i)$, conditionalize Equation 2 on i being present (implying $P(i) = 1$ in this new sample space), yielding

$$P(e \mid i) = p_i + P(a \mid i) \cdot p_a - p_i \cdot P(a \mid i) \cdot p_a. \quad (3)$$

That is, when i is present, e is the union of two independent events: e produced by i and e produced by a when i is present. Next, to explain $P(e \mid \bar{i})$, conditionalize Equation 2 on i being absent (implying $P(i) = 0$ in this sample space), yielding

$$P(e \mid \bar{i}) = P(a \mid \bar{i}) \cdot p_a. \quad (4)$$

That is, when i is absent, e is attributable to a alone.

The explanation of the two components of ΔP_i in Equations 3 and 4 is the heart of the power PC theory for generative causes. Reassembling these components of ΔP_i gives $\Delta P_i = P(e \mid i) - P(e \mid \bar{i}) = p_i + P(a \mid i) \cdot p_a - p_i \cdot P(a \mid i) \cdot p_a - P(a \mid \bar{i}) \cdot p_a$. Simplifying, we obtain

$$\Delta P_i = [1 - P(a \mid i) \cdot p_a] \cdot p_i + [P(a \mid i) - P(a \mid \bar{i})] \cdot p_a. \quad (5)$$

The rest of the power PC theory for generative causes consists of nothing but the mathematical consequences of Equation 5.

This equation summarizes one of the two relations between the probabilistic contrast model and the power PC theory, the one for the evaluation of generative power. This relation may be interpreted in two ways. First, just as the kinetic theory of gases explains how the movement of the theoretical particles produces the observable gas laws, Equation 5 explains how the theoretical entities p_i and p_a on the right-hand side (RHS) produce the observable ΔP_i on the left-hand side (LHS). Second, this equation expresses the difference between the probabilistic contrast model (as a model without a theory) and the power PC theory (a theory of that model). For a given focal set, the probabilistic contrast model bases its predictions on ΔP_i. In contrast, the power PC theory bases its predictions on p_i, which in general only partly determines ΔP_i. As becomes clear later when I consider the mathematical consequences of this equation, ΔP_i and p_i bear different relations to each other under different conditions.

Why Covariation Does Not, in General, Imply Causation

If an alternative cause a does exist (i.e., $p_a > 0$), and it does not occur independently of i (i.e., $P(a \mid i) \neq P(a \mid i)$), the RHS of Equation 5 has both a positive and a negative term in addition to the term containing p_i. It follows that ΔP_i is not interpretable as an estimate of p_i: It could overestimate p_i or underestimate it, depending on the values of $P(a \mid i)$ and $P(a \mid \bar{i})$. Specifically, if the reasoner believes that it is possible for a to occur more often in the presence of i than in its absence, he or she would think that this variation in how frequently a occurs might produce the positive contrast for i. For example, suppose a reasoner is trying to determine what causes mothers of young infants to be absent-minded. The reasoner observes a positive contrast for breast-feeding, his or her candidate cause: Proportionately more breast-feeding than non-breast-feeding mothers are absent-minded. The reasoner understands, however, that breast-feeding mothers might be more sleep deprived than their non-breast-feeding counterparts; they are more likely to be the person getting up at night to feed their infants. Suppose the reasoner understands that sleep deprivation is an alternative cause of absent-mindedness. According to Equation 5, the higher frequency of this alternative cause when mothers breast-feed their infants than when they do not can produce a positive contrast for breast-feeding, even if breast-feeding in fact does not cause absent-mindedness. The reasoner would therefore refrain from interpreting the positive contrast for breast-feeding as indicating its causal power. In sum, this equation shows that covariation does not, in general, imply causation.

Now, returning to show how one can evaluate whether i produces e (i.e., how p_i can be estimated), I rearrange Equation 5 to put p_i on the LHS, obtaining

[4] The potential influences of i and a on e are both direct in Pearl's (1988) terms.

$$p_i = \frac{\Delta P_i - [P(a|i) - P(a|\bar{i})] \cdot p_a}{1 - P(a|i) \cdot p_a}. \qquad (6)$$

When a Occurs Independently of i

Although ΔP_i, as has been shown, does not in general provide an estimate of p_i, there are conditions under which it can. Consider the special case in which a occurs independently of i, (i.e., $P(a|i) = P(a|\bar{i}) = P(a)$).[5] In this case, Equation 6 simplifies to

$$p_i = \frac{\Delta P_i}{1 - P(a) \cdot p_a}. \qquad (7)$$

Equation 7 indicates when and how well ΔP_i gives an estimate of p_i. It might be objected that, in Equation 7, one needs to know about p_a, a theoretical entity, to obtain an estimate of p_i. In cases in which the reasoner has prior knowledge about p_a, this knowledge can be applied. In other cases, however, it might seem that the problem of how causal inference begins reappears: The estimation of p_i begs the question of how p_a is estimated. The key to this problem offered by the power PC theory is that an estimation of p_a per se is not required. What is required is an estimation of the product $P(a) \cdot p_a$ (see Equation 7). This product yields the probability of the effect attributable to a within the focal set, a probability that can be estimated by observing the frequency of the effect in the absence of i, because a alone is present then (cf. Equation 4). (Because a occurs independently of i, the same estimate holds in the presence of i as in its absence.) This theory thereby circumvents the apparent circularity of needing to know about p_a.

Replacing $P(a) \cdot p_a$ in Equation 7 with its estimate, $P(e|\bar{i})$, yields

$$p_i = \frac{\Delta P_i}{1 - P(e|\bar{i})}. \qquad (8)$$

First, consider the extreme case in which $P(e|\bar{i}) \cong 0$ (i.e., the effect [almost] never occurs when the candidate is absent, for example, when alternative causes are constantly absent in the focal set). In this case, $p_i \cong \Delta P_i$, which means that in this optimal situation, the reasoner can interpret the contrast for i as a close estimate of the causal power of i.

Now, consider the other extreme case of Equation 8, in which $P(e|\bar{i}) \cong 1$. This is the situation in which the effect is (almost) always occurring, even when the candidate i is absent. In this case, p_i is undefined. In other words, $\Delta P_i \cong 0$ regardless of the magnitude of p_i, which means that, in this situation, the reasoner can no longer interpret the contrast for i as an estimate of the causal power of i.

Between the two extreme cases, as $P(e|\bar{i})$ increases from 0 to 1, when $\Delta P_i > 0$, p_i is increasingly larger than ΔP_i because it is equal to ΔP_i divided by an increasingly smaller number less than 1; in other words, ΔP_i is increasingly a conservative estimate of p_i. Two implications follow for situations under which a occurs independently of i. First, consider a reasoner whose goal is to judge whether a candidate factor is causal. For the sake of offering simple explanations, it seems that unless there is good evidence supporting the hypothesis that a candidate is causal, the null hypothesis that it is not should be the default. A conservative criterion errs on the side of promoting simple explanations. ΔP_i, being a generally conservative estimate, should therefore be regarded as a generally useful criterion for judging causation. Second, consider a reasoner whose goal is not simply to judge whether or not a candidate is a generative cause but also to estimate its causal strength. According to Equation 8, as $P(e|\bar{i})$ increases, a positive ΔP_i of the same magnitude will yield higher values of p_i, that is, higher estimates of the power of i. When $\Delta P_i = 0$, however, because $p_i = 0$ as long as $P(e|\bar{i}) < 1$, i should be judged noncausal whenever $P(e|\bar{i}) < 1$.

Relation to Experimental Design and Its Everyday Analogues

Although some of the mathematical consequences of Equation 5 correspond to apparent biases (as shown later), others are recognizable as principles of experimental design. I have shown that one consequence of this equation is that covariation does imply causation when alternative causes are believed to occur independently of the candidate (e.g., when alternative causes are constant) and $P(e|\bar{i})$ is not close to 1. Holding alternative causes constant while comparing conditions is, of course, a key principle of experimental design. Another consequence of Equation 5 is that a contrast of zero does not indicate that the candidate is noncausal when alternative causes are constantly producing the effect. This consequence corresponds to the concept of a ceiling effect.

Situations analogous to experiments sometimes occur naturally. Let me give an example. A toddler drops a ceramic item on the floor (a candidate cause), and it breaks (the effect). The contrast for dropping the item is therefore positive: Given dropping, the item breaks, but before the dropping (i.e., in its absence), the item does not break. The child might believe (as would an adult) that, except for the candidate cause and the effect, conditions before and during dropping remain unchanged. This belief is especially plausible when the sequence of events can be repeated at will. For this child, then, an episode of dropping a ceramic item forms a natural experiment. According to the power PC theory, the child should therefore be willing to infer from the positive contrast that dropping a ceramic item on the floor causes the item to break. Thus, the power PC theory explains the conditions under which covariation implies causation, in both experiments and their everyday analogues.

[5] First, note that this is not the same assumption as i and a independently producing e, which means that i produces e with the same probability regardless of whether a produces e. Second, note that Feller (1950/1957) defined two events as independent if their intersection has a probability equal to the product of the probabilities of the individual events. By this definition, independence between two events is symmetrical. Therefore, when a occurs independently of i, a and i occur independently of each other. I prefer the asymmetrical wording here, however, because it conveys more intuitively the idea that regardless of whether i is present, a occurs with the same probability, an idea that corresponds to the principle of keeping alternative causes constant in experimental designs.

Summary

The power PC theory not only specifies when covariation reveals causation, but also explains why covariation sometimes reveals causation and other times not. When a, the composite of alternative causes, does not occur independently of candidate cause i, ΔP_i does not reflect causal status. But when a does occur independently of i (e.g., when a is constant in a focal set), then to assess the generative nature of i, we see that, excluding the extreme case in which the effect (almost) always occurs (i.e., $P(e) \cong 1$) in the focal set, ΔP_i should provide an estimate of the causal status of i. When the effect (almost) never occurs in the absence of i (i.e., $P(e|\bar{i}) \cong 0$, e.g., when a is constantly absent), ΔP_i gives the closest estimate of the power of i. This estimate is increasingly conservative as $P(e|\bar{i})$ increases.

Main-Effect Contrast and Preventive Causal Power

Rather than raising the probability of an effect, some causes lower this probability. Such causes are often called preventive or inhibitory causes (e.g., Kelley, 1967, 1973). I assume that a preventive cause i has the power to stop an (otherwise occurring) effect e from occurring with probability p_i. To evaluate whether i prevents e, I make assumptions that are otherwise identical to those for evaluating generative causal power (see list of assumptions on p. 373).

I now explain how p_i—the preventive power of i—and p_a explain a nonpositive ΔP_i. (As before, a represents a composite of alternative causes that has a net generative effect.) In this case, the reasoner theorizes that the event "e occurring" is the intersection of two independent events: e produced by a and e not stopped by i. It follows that $P(e)$, the probability of the intersection, is the product of the probabilities of the constituent events:

$$P(e) = P(a)\cdot p_a\cdot[1 - P(i)\cdot p_i]. \qquad (9)$$

Using this theory to explain $P(e|i)$, I conditionalize Equation 9 on i being present, yielding

$$P(e|i) = P(a|i)\cdot p_a\cdot(1 - p_i). \qquad (10)$$

Using this theory to explain $P(e|\bar{i})$, I conditionalize Equation 9 on i being absent, yielding the same result as Equation 4: $P(e|\bar{i}) = P(a|\bar{i})\cdot p_a$. The explanation of the two components of ΔP_i in Equations 4 and 10 forms the heart of my theory for preventive causes. Replacing these components in ΔP_i with their explanations, it can be seen that

$$\Delta P_i = P(a|i)\cdot p_a - P(a|i)\cdot p_a\cdot p_i - P(a|\bar{i})\cdot p_a. \qquad (11)$$

The relation between the probabilistic contrast model and the power PC theory for the evaluation of generative causal power was shown earlier. Equation 11 shows the analogous relation between ΔP_i and p_i for the evaluation of preventive causal power.

Why Covariation Does Not, in General, Imply Causation

Equation 11 shows that, as for generative causes, if a does not occur independently of i, ΔP_i does not provide an estimate of p_i. Because the RHS of this equation has both a positive term and a negative term in addition to the term containing p_i, ΔP_i can overestimate p_i, or underestimate it, depending on the values of $P(a|i)$ and $P(a|\bar{i})$. For example, suppose that $p_i = 0$. A negative ΔP_i could result if a occurs less often in the presence of i than in its absence. In this case, ΔP_i overestimates the preventive power of i. Conversely, suppose that $p_i > 0$. If a occurs more often in the presence of i than in its absence, ΔP_i can underestimate the preventive power of i.

To see the power PC theory's predictions for p_i, I rearrange the terms in Equation 11 to put p_i on the LHS, yielding

$$p_i = \frac{[P(a|i) - P(a|\bar{i})]\cdot p_a - \Delta P_i}{P(a|i)\cdot p_a}. \qquad (12)$$

When a Occurs Independently of i

As for generative causes, although ΔP_i does not, in general, provide an estimate of the preventive power of i, there are conditions under which it can do so. Consider the case in which a occurs independently of i. Because $P(a|i) = P(a|\bar{i}) = P(a)$, Equation 12 simplifies to

$$p_i = \frac{-\Delta P_i}{P(a)\cdot p_a}. \qquad (13)$$

Recall that $P(a)\cdot p_a$ yields the probability of e attributable to a within the focal set, which can be directly estimated when a occurs independently of i by observing the frequency of e in the absence of i. Replacing $P(a)\cdot p_a$ in Equation 13 with its estimate, $P(e|\bar{i})$, yields

$$p_i = \frac{-\Delta P_i}{P(e|\bar{i})}. \qquad (14)$$

From Equation 14, it can be seen that in the one extreme in which $P(e|\bar{i}) \cong 1$, $p_i \cong -\Delta P_i$. That is, for reasoners who believe that a occurs independently of i, if they observe that e always occurs in the absence of i, they would regard ΔP_i as a good estimate of $-p_i$. This implies that if ΔP_i is a negative number, $-x$, then in this situation i stops the effect from occurring with a power of magnitude x. Note that this optimal situation for evaluating preventive causal power is opposite to that for generative causal power, in which alternative causes are constantly absent.

In the other extreme in which $P(e|\bar{i}) \cong 0$ (because $P(a) \cong 0$ or $p_a \cong 0$), it can be seen that p_i is undefined. In this case, $\Delta P_i \cong 0$ regardless of the value of p_i. That is, if the reasoner observes that e never occurs in the absence of i, she or he would believe that it is not possible to estimate p_i from ΔP_i. Note that the values of $P(e|\bar{i})$ for which p_i is undefined are at opposite extremes for the evaluation of generative and preventive causal powers.

In between these two extremes, when $0 < P(e|\bar{i}) < 1$, the reasoner should regard a negative ΔP_i as a conservative estimate of $-p_i$. Moreover, for a negative ΔP_i of the same magnitude, as $P(e|\bar{i})$ decreases, higher strengths should be inferred for i. This direction of change relative to the value of $P(e|\bar{i})$ is opposite to that in the interpretation of a positive contrast. When

$\Delta P_i = 0$, however, $p_i = 0$, according to Equation 14, as long as $P(e|\bar{i}) > 0$.

Summary

Analogous to my analysis of a positive contrast, the power PC theory explains as well as specifies the conditions under which negative covariation reveals the operation of a preventive cause. In this case, when the composite alternative cause occurs independently of inhibitory candidate cause i, excluding the extreme case in which the effect (almost) never occurs in the focal set, ΔP_i should provide an estimate of the preventive power of i. When the effect (almost) always occurs in the absence of i, ΔP_i gives the closest estimate of the power of i. This estimate is increasingly conservative as $P(e|\bar{i})$ decreases. This direction of change in the estimation of a negative ΔP_i relative to the value of $P(e|\bar{i})$ is opposite to that in the interpretation of a positive ΔP_i. As for generative causes, however, when inhibitory candidate i does not occur independently of a, ΔP_i does not reflect causal status.

Predictions of the Power PC Theory

The two theoretical interpretations of the probabilistic contrast model in terms of generative and preventive causal powers specify when covariation does and does not reveal these kinds of powers. These interpretations also explain why it does when it does. My analysis generates many testable predictions. As mentioned earlier, I propose that people implicitly follow a qualitative version of my mathematical analysis rather than the analysis itself. The predictions of the power PC theory are therefore ordinal.

One prediction is that even untutored reasoners, including children, will be unwilling to judge a covarying candidate to be a genuine cause (to use Suppes's, 1970, terminology) if they *are aware of the existence of an alternative cause and believe* that it does not occur independently of the candidate cause in their focal set. That is, reasoners have an implicit understanding of potential confounding by the alternative causes that they know about. (This is not the same as understanding possible confounding by unknown alternative causes, which would require explicit knowledge of the variables in the power PC theory, independent of their instantiations.) If they misjudge a spurious cause to be a genuine one (i.e., erroneously infer causality from covariation), it is because they erroneously believe that the conditions for estimating causal power are met when they are not. For example, a reasoner may believe that a occurs independently of i when it actually does not. Likewise, she or he may believe that there are no alternative causes of an effect when there are. In these situations, the reasoner would erroneously interpret ΔP_i as an estimate of p_i.

Thus, one boundary condition for estimating causal power from covariation is that alternative causes occur independently of the candidate cause. For situations in which reasoners believe that this condition is met, the power PC theory predicts that they would judge a candidate with a noticeably nonzero contrast to be a genuine cause. Even for such situations, however, this theory predicts additional boundary conditions for the interpretability of a contrast of zero. According to Equations 7 and 13, whether such a contrast is interpretable as an estimate of causal power depends on two factors. One is the base rate of the effect (i.e., $P(a) \cdot p_a$, which is equal to $P(e|\bar{i})$). When this rate is clearly between 0 and 1, a noncontingent candidate should be judged as noncausal. When this rate approximates either 0 or 1, however, judgment depends on the goal of the inference: Is one concerned with assessing the generative or preventive nature of the candidate? Just as the kinetic theory of gases explains the boundary conditions for Boyle's law, the power PC theory explains the boundary conditions for contrast, including the diametrically opposite conditions that depend on the type of power being assessed.

In turn, the interaction between the base rate of the effect and the nature of the assessment for noncontingent candidates explains and predicts the variations in causal judgments regarding such candidates noted in the introduction. The doctor in my food-allergy anecdote could not tell whether a food item caused hives because scratching, an alternative cause, was always present and always producing hives, in which case the denominator in Equation 7— $1 - P(a) \cdot p_a$ —is 0. Analogously, the researcher in my headache anecdote could not tell whether the drug relieves headaches because the test is conducted among a population that does not have headaches, in which case the denominator in Equation 13—$P(a) \cdot p_a$—is 0. These are the two cases in which ΔP_i does not give an estimate of p_i even when alternative causes are controlled. In addition, Equation 7 explains the difference in the certainty of causal judgments for potentially generative candidates with a zero contrast depending on whether the effect always or only sometimes occurs.

In addition to specifying boundary conditions for interpreting contrasts, the power PC theory also makes predictions about changes in estimated causal strength for candidates with the same contrast as the base rate of the effect changes. For candidates with a positive ΔP of the same magnitude, as the base rate of the effect increases, the candidate will be inferred to *have a greater causal power. An opposite trend is predicted for* negative ΔPs of the same magnitude as a function of this base rate. A less obvious prediction of the power PC theory (derived in a later section) is that reasoners should weigh frequencies of the effect in the presence of a candidate cause more than those in its absence. Both of these predictions have been regarded as biases.

Seeking Causation From Covariation

Note that, according to my analysis, to infer causation from covariation the reasoner need not know what the alternative causes are (i.e., the identity of a) or how strong they are (i.e., the value of p_a). Such information can be used when it is available (Equations 7 and 13), but it is not required (Equations 8 and 14). Rather, the reasoner needs only to know that the alternatives (whatever they are) occur independently of the candidate cause (e.g., are constant) and to observe the base rate of the effect in the relevant context, noting the extreme base rates that disallow causal inferences. The a priori knowledge required is that there are such things as causes that have the power to produce an effect or to inhibit it. This knowledge, which is embodied in the process of induction, enters my analysis in the form of variables (i.e., p_i and p_a in Equations 5 and 11) that do not require any prespecified values, variables that may adopt a causal or a noncausal value depending on purely observable events. The a priori knowledge is therefore domain independent. Because the reasoner need not know the identity or the magnitude of any cause before inducing the causality of a candidate cause, there is no circularity in my analysis. This analysis explains how causal inferences can begin: Given the a priori knowledge, inference of causality from covariation is justified under a specified set of conditions.

The analysis I presented specifies how reasoners would interpret a given ΔP under various conditions. This analysis does not directly specify which conditions a reasoner would seek among those possible. For reasoners whose goal is to infer causal power, however, it seems plausible that they should attempt to obtain the best estimates of the causal powers of candidate causes. My analysis justifies their choices. A comparison between Equations 6 and 7 justifies why such a reasoner should prefer to assess covariation in focal sets in which alternative causes are constant: a occurs independently of candidate cause i in these focal sets. The optimal way to obtain a set of events in which a occurs independently of i is to manipulate i. If manipulation is impossible, reasoners might attempt to select sets of events among those observed in which a occurs independently of i, although they are likely to be less confident that independence holds. Likewise, Equations 7 and 13 justify why reasoners should prefer focal sets in which alternative causes are constantly absent to assess the generative nature of a candidate cause, whereas they should prefer those in which alternative causes are constantly present to assess the preventive nature of a candidate cause: Contrast gives the closest estimate of causal power in such focal sets.

Even among focal sets in which a occurs independently of i, however, not all will allow an estimation of the causal power of i. If there is one set, and only one set, that does allow this estimation, there would be no conflicting information, and reasoners should adopt that set as their focal set. If there is more than one available set that reveals causal power, but these sets are consistent in the causal power they indicate, there would still be no conflicting information. If the causal powers revealed in multiple informative sets conflict, however, reasoners would have to either withhold judgment or resolve the conflict in some way. Finally, if none of the focal sets in which a occurs independently of i allows an unambiguous estimation of causal power, or if the information available does not allow any partitioning that renders a independent of i, reasoners would have to either withhold judgment or select the next best available set (or sets) with reduced confidence if forced to make a decision.

Once the leap from observed covariation to underlying causal power has been made, people may apply their acquired causal knowledge to causal judgments about cause–effect relations that are believed to be of the same kind (Cheng & Lien, 1995). For example, learning that remote controls can operate a television set and a driveway gate "prepares" the reasoner to accept that the covariation between pressing a remote control and the opening of a garage door is also due to a similar causal power. The generalization of acquired causal knowledge extends the scope of a covariational relation that has been deemed causal.

Some researchers have pitted covariational models of the induction process against evidence for the influence of prior domain-specific causal knowledge, suggesting that the latter process offers an alternative solution to the problem of when covariation implies causation (e.g., Ahn & Bailenson, 1996; Ahn et al., 1995; Shultz, 1982; White, 1989; see Cheng, 1993, for an analysis). Ahn et al. (1995), for example, wrote that "people seek out and prefer information about causal mechanisms rather than information about covariation" (p. 299) and that "when direct information about mechanisms is difficult or impossible to attain, . . . covariation methods can be useful heuristics" (pp. 339–340) for distinguishing between "true causality and spurious correlation" (p. 341). Work adopting this approach has demonstrated how prevalent the influence of prior domain-specific causal knowledge is, convincingly arguing that the problem of when covariation implies causation in many cases should be pushed one step back in time, to the acquisition of the causal mechanisms. As mentioned earlier, however, this approach ultimately fails to provide an answer to that problem.

A Computational-Level Analysis of the R–W Model

A goal of this article is to analyze the R–W model in terms of causal power. A causal power analysis of a model, however, requires as a prerequisite a mathematical function characterizing the model's asymptotic behavior. Connectionist models, which specify an algorithm and a representation, typically do not permit such characterizations. To make predictions for these models, researchers generally have to rely instead on computer or thought simulations of specific experiments. One of the attractions of the R–W model is that it turns out to be an exception to this rule. I now present an analysis of what function the R–W algorithm asymptotically computes.

The R–W model represents the learning of an association between cue i (e.g., a tone that is present in the current event) and outcome j (e.g., shock) by a change in the strength of a link between two elemental units in a network, one representing cue i and the other representing outcome j. (Cue i and outcome j are traditionally termed the *conditioned stimulus* and the *unconditioned stimulus,* respectively. In causal terms, if each cue i is a candidate cause, then j is the effect.) For any cue i that is present during the event, strength is revised according to the rule

$$\Delta V_{ij} = \alpha_i \beta_j \left(\lambda_j - \sum_{k=1}^{n} V_{kj} \right), \tag{15}$$

where ΔV_{ij} is the change in associative strength between cue unit i and outcome unit j as a result of the current event, α_i and β_j are rate parameters that respectively depend on the salience of i and j, and λ_j is the actual outcome. Typically, if the outcome is present, λ_j is defined as 1; if the outcome is absent, this value is defined as 0. Similarly, β_j is typically assumed to be a larger number when the outcome is present than when it is absent. $\Sigma_{k=1}^{n} V_{kj}$, the outcome predicted by the model, is defined as the sum of the current strengths of links to unit j from all units representing the n cues present in that event. If cue i is absent during the event, the associative strength of its cue unit remains unchanged. (This restriction of strength revision to present cues implies that α_i has a weight of 0 when i is absent.) Learning continues until there is no discrepancy between the actual and predicted outcomes (averaged over a number of trials). The strengths that are updated according to Equation 15 are equivalent to weights on the links in a two-layered connectionist network, with the predicting cues being represented on the input layer and the predicted outcome on the output layer.

Asymptotic Weights for Designs for Which the R–W Model Does and Does Not Compute Conditional Contrasts

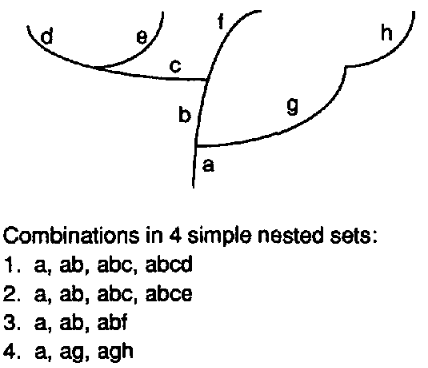
In Appendix A, I show that, assuming that (a) λ equals 1 for trials on which the effect occurs and 0 otherwise and (b) the learning rates remain constant across trials on which the effect does and does not occur, for experimental designs that satisfy a condition I term *nesting,* the R–W model asymptotically computes conditional contrasts. The first assumption allows the strengths of connections to be interpreted as probabilities when the nesting condition is satisfied. The parameter λ is a scaling factor in this model (see Rescorla & Wagner, 1972). The second assumption means that both learning rates, α_i and β_j, are constant across trials on which the effect does and does not occur for all trials relevant to the strength of i. (Recall that the strength of a cue is not updated when it is absent; therefore, i is present on all of these trials.) Although α_i, being associated with the cue rather than the effect, is typically assumed to be constant for a given cue across these trials, β_j is often assumed to be greater when the effect occurs than when it does not. Although many researchers conduct simulations of the R–W model under this assumption, most asymptotic predictions of this model, including blocking, conditioned inhibition, extinction of conditioned inhibition, superconditioning, and overexpectation, are in fact independent of this assumption (see Miller et al., 1995; see also the derivations of some of these predictions in Melz et al., 1993). The two well-known exceptions concern the phenomena of the relative validity of cues (Shanks, 1991; Wagner et al., 1968; Wasserman, 1990) and the influence of the base rate of the effect on the perception of contingency (Wasserman et al., 1993). I separately discuss these special cases. For all other cases, my derivation directly applies.

I first show that in a design with multiple cues, if every combination of cues in the design except the one with a single cue can be characterized as a proper superset of all combinations with fewer cues, then the design is nested. For example, the design with cue combinations a, ab, and abc is nested, whereas the design with cue combinations a, ab, and bc is not nested. (A letter denotes a cue, and a cluster of letters denotes a cue combination; for example, ab denotes the combination consisting of cues a and b.) According to the R–W model, for any combination with multiple cues in a nested design, the sum of the strengths of the cues in it that do not belong to the next smaller combination is asymptotically equal to the contrast for those cues (as a composite) conditional on the presence of the cues in the smaller combination (i.e., the rest of the cues in the larger combination; Equation A23 in Appendix A). For example, in the design with the combinations a, ab, and abc, the strength of c is equal to the contrast for c conditional on the presence of both a and b, and the strength of b is equal to the contrast for b conditional on the presence of a. I also show in Appendix A that when the cue combinations are not nested, the strength of a cue is not, in general, equal to any of its (conditional or unconditional) contrasts.

I then generalize my definition of nesting to designs involving partially overlapping cue combinations. I refer to the type of nesting that does not involve such combinations (which I just described) as simple nesting. By partially overlapping, I mean that the combinations are neither disjoint nor a superset or subset of each other. I show that, for such designs, the R–W model still asymptotically computes conditional contrasts as just specified if, for every pair of partially overlapping combinations, all supersets of one combination (including the combination itself) share the same intersection with the other combination, and this intersection occurs as a separate combination. An example of this type of nesting is the design a, ab, and ac. The intersection, a, of the partially overlapping sets ab and ac occurs as a separate combination. Therefore, the strength of c is equal to the contrast for c conditional on the presence of a; likewise, the strength of b is equal to the contrast for b conditional on the presence of a. The example of an unnested design given earlier is still unnested by this definition because the intersection, b, of the partially overlapping sets ab and bc does not occur as a separate combination.

A visual characterization of a nested set that contains partially overlapping sets is that the partially overlapping sets form a "tree" structure in which there are multiple "branches," with each path from the bottom of the trunk to the tip of a branch forming a simple nested set. This characterization assumes that cues along the same path below a cue are cumulative, as illustrated in Figure 1. In this figure, each branch of the tree accompanied by a label represents a stimulus. For an example of a simple nested set within this tree, consider the path from the trunk (which happens to have only one cue) labeled a to the tip of the branch labeled d. The simple nested set represented by this path consists of the following combinations: a, ab, abc, and $abcd$. The figure shows partially overlapping sets in which the branching is upward, as in a tree, as my definition of nesting implies. A set of combinations would be unnested if the analogous representation does not form a possible tree, in that cues along the same path are not cumulative as defined or the branching occurs downward, as when trunks from multiple trees grow together or when the tips of different branches merge.

In summary, *in a design with multiple cues, if there are no*



Combinations in 4 simple nested sets:
1. a, ab, abc, abcd
2. a, ab, abc, abce
3. a, ab, abf
4. a, ag, agh

Figure 1. Tree representation of a nested set that contains partially overlapping stimulus combinations.

partially overlapping cue combinations unless for every pair of such combinations, (a) all supersets of one combination share the same intersection with the other combination, and (b) this intersection occurs as a separate combination, then the design is nested. In other words, except for such partially overlapping combinations, every combination of stimuli in a nested design can be characterized as a proper superset of any combination that contains some but not all stimuli in it. *According to the R–W model, for any cue combination in a nested design, the total strength of the stimuli in it that do not belong to the next proper subset is asymptotically equal to the contrast for those stimuli (as a composite) conditional on the presence of the stimuli in the smaller combination (i.e., the rest of the stimuli in the larger combination).* When a design is not nested, except for special cases, the R–W model does not compute conditional contrast.

Application of the Derivation to Experimental Paradigms

A mathematical analysis of the asymptotic behavior of the R–W model yields an understanding of what the associative strengths computed by the model actually represent. This understanding is useful in several ways. First, it eliminates the need to conduct computer simulations of the R–W model for many experimental designs. More important, this understanding eliminates uncertainty regarding differences in predictions between this and other models. When two models make the same prediction for a given situation, it is possible that they actually compute the same function or that they compute different functions that happen to yield the same value. Only in the latter case should an investigator look for variations for which the models yield different values using the same design. The application of my analysis to experimental designs allows one to deduce when each of these two possible relations is true. My analysis specifies (a) the conditions under which the R–W model does and does not compute conditional contrasts and (b) the contrast it computes when it does compute a contrast. For some of these designs, the R–W model asymptotically computes the exact same function as another model, that of Cheng and Holyoak (1995). Most important, my analysis enables an interpretation of the R–W model in terms of causal power— the determinant of whether a model's prediction will generalize beyond the set of data on which it is based. As I show later, although the R–W model asymptotically computes conditional contrasts when the cue combinations have a nested structure, not all such contrasts computed by the model provide an estimate of causal power. When this model computes a conditional contrast that gives an estimate of causal power, it accounts for the observed results; otherwise, it does not. An interpretation in terms of causal power pinpoints how and why the R–W model fails when it fails and, hence, might guide the development of a superior algorithmic-level model.

In the next section, I apply the result of my derivation to find the contrast computed by the R–W model for many well-known designs or adaptations of them in the classical conditioning literature: unconditional contingency (e.g., Rescorla, 1968; Wasserman et al., 1993), blocking (e.g., Chapman & Robbins, 1990; Fratianne & Cheng, 1995; Kamin, 1968; Shanks, 1991), induced overshadowing (e.g., Baker et al., 1993; Price & Yates, 1993), overexpectation (e.g., Park & Cheng, 1995; Rescorla, 1970), acquisition of conditioned inhibition (e.g., Miller & Schachtman, 1985; Williams, 1995; Williams & Docking, 1995; Yarlas et al., 1995), and extinction of conditioned inhibition (e.g., Hallam, Matzel, Sloat, & Miller, 1990; Miller & Schachtman, 1985; Williams & Docking, 1995; Yarlas et al., 1995; Zimmer-Hart & Rescorla, 1974). My analysis applies to these designs because the asymptotic predictions of the R–W model for them are not dependent on a difference between learning rates when the outcome does and does not occur.

Before considering specific experimental designs, however, let me note three phenomena the explanations of which are beyond the reach of the R–W model. This model fails in these cases because its predictions are based on a single output parameter: the strength of the association between a stimulus and the outcome. First, this feature of the model renders it incapable of explaining why people can be simultaneously aware that a relation is covariational (i.e., $\Delta P \neq 0$) and that it is noncausal (e.g., dinner covaries with sunset but does not cause it). In contrast, the power PC theory has two output parameters: the output of the probabilistic contrast model, ΔP, and that of the theory of this model, p_i. Thus, Equations 5 and 11, respectively, explain how a positive ΔP and a negative one can fail to be interpretable estimates of causal power. Second, the R–W model cannot explain the distinction people make among a cause, an enabling condition, a causally irrelevant factor, and a novel factor. An enabling condition is not simply a cause with an intermediate strength, and a novel factor is neither a factor with zero strength nor one with an intermediate strength. Finally, the R–W model cannot explain what my allergy and headache anecdotes illustrate: Even when alternative causes occur independently of the candidate (so that ΔP is not confounded), whereas some zero contrasts indicate noncausality, others indicate that no causal inference can be drawn. Experimental evidence for the third phenomenon is included in the following section.

Empirical Tests of the Power PC Theory and the R–W Model

Reasoners can potentially compute contrasts over indefinitely many possible partitions of events. Do they indeed select sets of events among those available that optimally reveal causal power? For example, do they attempt to select focal sets in which alternative causes occur independently of the candidate cause? For such sets, do reasoners interpret nonzero contrasts according to the power PC theory, that is, as a joint function of conditional ΔP_i and $P(e|\bar{i})$, as specified in Equations 8 and 14? Even among such sets, not all zero contrasts reveal causal power. Do reasoners interpret zero contrasts according to whether they reveal causal power? I now describe some empirical tests of the power PC theory that address these questions. I first review tests of the selection of focal sets and the interpretation of contrasts as estimates of causal power in studies involving multiple varying candidate causes. These include tests of the interactive predictions regarding the boundary conditions for interpreting contrasts as estimates of generative and preventive causal power. I then review studies involving a single varying candidate. These studies test the power PC theory's prediction that the magnitude of causal estimates is a joint function of ΔP_i and $P(e|\bar{i})$.

In this section, in addition to evaluating the power PC theory against these findings, I evaluate two competing accounts of causal induction against them: the R–W model and the traditional contingency model. I also discuss several studies that have been interpreted as contradicting the probabilistic contrast model and, hence, might be interpreted as contradicting the power PC theory.

Selection of Focal Sets: Tests of the Traditional Contingency Model Involving Multiple Varying Candidate Causes

One answer to the question of whether reasoners attempt to select focal sets in which alternative causes occur independently of the candidate cause comes from studies testing the traditional contingency model. Before Cheng and Novick's (1992) proposal that reasoners select sets of events for computing contrast, all causal induction models in psychology assumed, by default, that whatever function is computed is computed over the universal set of events, which, for an experiment using unfamiliar materials, implies the entire set of events in the experiment. For the traditional contingency model, this means that unconditional contingency is what the reasoner is assumed to compute (e.g., Baker et al., 1989, 1993; Chapman & Robbins, 1990; Dickinson et al., 1984; Price & Yates, 1993; Rescorla, 1968; Shaklee & Tucker, 1980; Shanks, 1985a, 1985b, 1987, 1991; Ward & Jenkins, 1965; Wasserman et al., 1993). Some of these studies involved test situations that included multiple varying candidate causes. Equations 5 and 11 show that for such situations, the traditional contingency model does not, in general, provide estimates of causal power, because the contingency for a candidate cause in the universal set can be confounded by alternative causes. Do reasoners use the universal set (i.e., base their causal judgments on unconditional contingency), or do they select a set in which alternative causes are constant (i.e., base their causal judgments on conditional contingency), as predicted by the power PC theory?

Some researchers tested the traditional contingency theory against the R–W model (e.g., Baker et al., 1993; Chapman & Robbins, 1990; Price & Yates, 1993; Shanks, 1991). These researchers have interpreted their results as supporting the R–W model. All of these studies controlled for unconditional contrast across conditions but varied conditional contrast for the target candidates between conditions. As has been noted in several articles (Cheng & Holyoak, 1995; Melz et al., 1993; Shanks, 1995; Spellman, 1996b), support for the R–W model in these studies in fact implies support for the conditional contingency account. The R–W model asymptotically computes conditional contrast in all except one of these experiments (Shanks, 1991, Experiment 3, which I discuss separately). I illustrate subsequently how these studies also support the power PC theory.

Induced Overshadowing

Some of these studies adapted the overshadowing design from Pavlovian conditioning (Mackintosh, 1983), in which a stronger and a weaker cue that are presented in combination are introduced in the same phase and receive an equal amount of information. These studies differed from the traditional overshadowing design in that variations in the salience of the cues were induced during the experiment.[6] In these induced overshadowing studies (all 16 designs used in Baker et al., 1993, except the ones labeled PR.5/1, PR.5/−1, and PR−.5/1; Price & Yates, 1993; Spellman, 1996a), denoting the two varying cues as A and B and the context as C, the cue combinations are C, AC, BC, ABC. These cue combinations are not nested, because the tips of two branches (A and B) merge. The frequencies of the various trial types presented in these studies, however, happen to allow the R–W model to compute conditional contrasts. In Appendix B, I show that, for this design, the R–W model converges on the same solution as for a nest set (e.g., C, BC, and ABC) formed by ignoring a combination (e.g., AC) if the contrast for a varying cue (e.g., A) conditional on the presence of the other varying cue and the context (i.e., $P(e|ABC) - P(e|\bar{A}BC)$) is equal to its contrast conditional on the absence of the other cue and the presence of the context (i.e., $P(e|A\bar{B}C) - P(e|\bar{A}\bar{B}C)$). This condition holds for all of the induced overshadowing studies just mentioned. The R–W model therefore computes conditional contrasts in these studies just as in nested designs.

To interpret the results of these studies with respect to the power PC theory, I consider separately (a) comparisons between contingent candidate causes (i.e., candidates with nonzero *conditional* contingencies), in which one candidate had a higher conditional contrast than the other (Baker et al., 1993, Condition .5/.8 vs. Condition .5/0 in Experiment 2; Price & Yates, 1993; Spellman, 1996a), and (b) comparisons involving at least one noncontingent candidate cause (i.e., a candidate with a *conditional* contingency of 0; all other comparisons between induced overshadowing designs in Baker et al., 1993; Spellman, 1996a).

For the former type of comparisons, the conditional contrasts all satisfy the boundary conditions of the power PC theory (Equations 8 and 14 show that only zero contrasts can fall outside these boundaries). Because the R–W model computes conditional contrasts, its ordinal predictions often coincide with those of the power PC theory. They do for all of these comparisons.

Comparisons between contingent candidate causes. Spellman (1996a), for example, compared causal judgments regarding candidate causes that had the same unconditional contingency of zero; however, conditional on the absence of the alternative cause, the candidate in one condition had a contrast of .33, and the candidate in another condition had a contrast of −.33. Because one conditional contrast was positive and the other was negative, regardless of the values of $P(e|\bar{i})$ that enter into the computation of these conditional contrasts, the power PC theory predicts that the candidate with the more positive conditional contrast should receive a more positive causal rating than the other. This prediction was confirmed.

Likewise, participants in Baker et al. (1993) rated the target candidate cause (camouflage for a tank under their cover story) as more causal in the .5/0 condition than in the .5/.8 condition,

[6] The traditional overshadowing phenomenon can be explained by my theory only with the aid of auxiliary assumptions regarding the salience of the cues. Note that, to explain this phenomenon, the R-W model also requires assumptions regarding the salience of the cues.

even though unconditional contingency for the candidate cause was .5 in both conditions. Causal power can be estimated from either the focal set in which the alternative candidate (a plane in their cover story) was present or that in which it was absent. Assuming that participants based their causal judgments on the latter focal set, the one in which ΔP gives a better estimate of causal power, they would judge camouflage to be more causal in the former condition. In this condition, the contrast for camouflage conditional on the absence of the plane was .50, with $P(e|\bar{i}) = .25$ (where i represents camouflage), yielding a causal power of .67 according to Equation 8. In the latter condition, this contrast was .21, with $P(e|\bar{i}) = .04$, yielding a causal power of .22. The power PC theory therefore explains Baker et al.'s results. An analogous interpretation in terms of causal power applies to Price and Yates's (1993) results, which showed the same pattern as Baker et al.'s (see Spellman, 1996b, and Shanks, 1995, for detailed analyses of Price and Yates's experiments in terms of conditional contrasts).

In sum, all comparisons involving contingent candidates show that participants based their causal judgments on focal sets in which alternative causes were constant rather than on the universal set. These results support the power PC theory as well as the R–W model.

Comparisons involving a noncontingent candidate cause. Comparisons between contingent and noncontingent candidate causes always showed a higher absolute rating for noncontingent candidates, despite their identical unconditional contingencies (Conditions .5/1 vs. .5/0, −.5/−1 vs. −.5/0, .5/−1 vs. .5/0, −.5/1 vs. −.5/0, and PR.5/0 vs. PR.5/.4 in Baker et al., 1993; Spellman, 1996a). Because some of the zero contrasts were undefined for the evaluation of either generative or preventive causal power, the interpretation of this finding with respect to the power PC theory requires an assumption: When participants were uncertain of the causal status of a candidate that had a zero contrast, they rated it as less causal than a noncontingent candidate. This assumption is necessary as a result of two sources of ambiguity: (a) None of these studies provided a context that clearly indicated whether the participants were to evaluate generative or preventive causal power, and (b) none of the studies provided participants with the option of expressing uncertainty.

It might be argued that participants in the studies just discussed, which used a within-subject design, simply selected focal sets in which contrast varied across conditions, perhaps as a result of an experimental demand to give different answers across conditions. To rule out this explanation, Spellman compared causal judgments across conditions in which the conditional contrast for the candidate cause remained at 0 but the unconditional contrasts varied, from −.5 to 0 to .5. She reported that participants rated the candidate in these three conditions highly similarly, showing that they based their causal judgments on focal sets in which alternative causes were controlled, regardless of which type of contrast varied. Other studies comparing two or more noncontingent candidates, however, showed more variable ratings. For example, of the 12 conditions (some with the same design) in Baker et al. (1993) in which the target candidate was noncontingent, 3 showed a rating that reliably deviated from 0 (e.g., Conditions 0/1 and 0/0 in Experiment 1).

Comparisons between noncontingent candidates in these studies are difficult to interpret because of the two sources of ambiguity mentioned earlier; given that the kind of causal power to be assessed was not specified to the participants, whether a particular zero contrast is defined with respect to the power PC theory could depend on the nature of the assessment assumed by a participant. Moreover, even when such a contrast is defined, small misperceptions of ΔP can lead to large apparent biases for noncontingent candidates, as I explain when I review studies involving a single varying candidate. A study by Fratianne and Cheng (1995), described later, avoided these interpretation problems.

Blocking

Some studies testing the traditional contingency model used the blocking design borrowed from Pavlovian conditioning. In this design, the first phase presents trials pairing the presence of a predictive cue with an outcome and the absence of this cue with the absence of the outcome (e.g., Chapman & Robbins, 1990; Dickinson et al., 1984; Shanks, 1991, Experiment 2). The second phase presents trials pairing the combination of this predictive cue and a novel cue (the candidate cue in question) with the outcome. These cue combinations are nested; hence, the R–W model computes conditional contrasts. For example, in Chapman and Robbins's study (1990, Experiment 1), participants were presented trials pairing the presence of a predictive cue, P, with the outcome ($P+$). They were also presented with trials pairing the presence of a nonpredictive cue, N, with the absence of the outcome ($N-$). On trials in which no cue was present, the outcome was absent. Then, in a second phase, a novel cue B (the to-be-blocked cue), in combination with P, was paired with the outcome ($PB+$). Likewise, another novel cue C (the control cue), in combination with N, was paired with the outcome ($NC+$). Assuming the constant presence of a context cue, X, to represent trials on which no explicit cue is present, the design in this experiment was $X-$, $XP+$, $XPB+$, $XN-$, $XNC+$. The cue combinations X, XP, XPB, XN, and XNC are nested; there are partially overlapping combinations (XP and XPB each partially overlap with XN and XNC), but for every pair of partially overlapping combinations (e.g., XP and XN), all supersets of one combination (e.g., XP and XPB) share the same intersection with the other combination (they both intersect with XN by X), and this intersection (X) occurs as a separate combination. In other words, the paths from the respective "tips," B and C, of the two branches of the tree structure are the simple nested sets (a) X, XP, XPB and (b) X, XN, XNC. It follows that the strength of B according to the R–W model is equal to its contrast conditional on the presence of X and P, and this contrast is zero (because $P(e|XPB) = P(e|XP\bar{B}) = 1$. This model therefore predicts that the strength of B should asymptotically be zero. Similarly, it follows that the strength of C is equal to its contrast conditional on the presence of X and N, and this contrast is 1 (because $P(e|XNC) = 1$ but $P(e|XN\bar{C}) = 0$). This model therefore predicts that despite the same unconditional contingency for B and C, B should have a lower causal strength than C.

This ordinal prediction has been supported in many studies using variants of the blocking design (e.g., Chapman, 1991;

Chapman & Robbins, 1990; Dickinson & Burke, 1996; Dickinson et al., 1984; Shanks, 1991): The blocked cue received a lower causal rating than a control cue that has been identically paired with the outcome (and has the same unconditional contingency) but has been presented only in combination with (a) a nonpredictive cue (e.g., Chapman, 1991; Chapman & Robbins, 1990; Dickinson & Burke, 1996; Shanks, 1991, Experiment 2) or (b) another novel cue (e.g., Chapman, 1991; Dickinson et al., 1984). These findings, which have often been regarded as support for the R–W model, equally support an account in terms of conditional contrasts because the R–W model computes conditional contrasts in these studies (for similar analyses of variants of the blocking design, see Melz et al., 1993). These results show that, rather than using the universal set, reasoners select focal sets in which alternative causes are controlled.

Note, however, that the zero contrast for B is undefined according to Equation 8, because $P(e|X P\bar{B}) = 1$. If participants assumed that they were to evaluate generative causal power, the power PC theory would predict that participants should be uncertain of the causal status of B, in contrast to the R–W model, which predicts that participants should be confident that B is noncausal. All human studies using the blocking design in which (a) the trials were discrete (so that the power PC theory applies) and (b) the predictive (i.e., blocking) cue was always paired with the effect (so that the zero contrast for the to-be-blocked cue is undefined) showed a mean rating for the indeed blocked cue that was about midway between being a cause and a noncause (Chapman, 1991; Chapman & Robbins, 1990; Dickinson et al., 1984; Shanks, 1991; Shanks & Dickinson, 1987; Waldmann & Holyoak, 1992; Williams et al., 1994; it is not clear whether this was true of Dickinson & Burke's study, 1996, because they did not directly report ratings but, rather, reported differences between ratings). This finding of partial blocking has been interpreted to reflect uncertainty and, hence, to contradict the R–W model (Cheng & Holyoak, 1995; Melz et al., 1993). This interpretation is plausible given that an intermediate rating has often been reported for candidates about which participants received no information at all (e.g., Williams, 1995; Williams & Docking, 1995). To the contrary, Shanks (1993) interpreted the intermediate ratings to be consistent with the R–W model by arguing that performance in these studies was merely preasymptotic. It is not possible to disambiguate the interpretation of the intermediate ratings without further experimentation. First, there is no clear definition of when performance is asymptotic. More important, these studies—not having been designed to test the power PC theory—neither specified the type of causal power to be assessed nor allowed the option of expressing uncertainty.

This difference in the interpretation of a zero contrast points to an important theoretical distinction between the power PC theory and the R–W model. The R–W model bases its prediction regarding the causal ratings of the blocked cue directly on the (conditional) contrast for this cue. An explanation of this prediction was provided by Cheng and Novick (1992), who derived that, if one assumes that causes have the power to produce their effects with some positive probability, the contrast for a candidate cause with a given power is reduced, in the extreme to zero, when alternative independent causes of the same effect are present. That is, they explained the phenomenon of blocking in terms of this reduction in the magnitude of contrast. My current analysis turns this explanation on its head: Whereas Cheng and Novick assumed that causal judgments are directly based on contrast, the value of which they predicted on the basis of assumptions about causal power, I now assume that causal judgments are directly based on causal power, which is only indirectly estimated by contrast. This difference in the interpretation of the blocked cue's zero contrast should allow discrimination between the power PC theory and the R–W model if the dependent measure is sensitive to this difference.

Summary

Studies testing the traditional contingency model using multiple candidate causes uniformly show that, instead of using the universal set, reasoners select focal sets in which alternative causes occur independently of the candidate cause (i.e., sets that optimally reveal causal power) and make causal judgments according to contrasts in these sets. All interpretable results strongly support the power PC theory. These studies do not, however, discriminate the power PC theory from the R–W model because their procedures underspecified what is required (according to the power PC theory) for an unequivocal interpretation of observed causal judgments.

Selection of Focal Sets: Tests Discriminating Between the Power PC Theory and the R–W Model

Some studies that test the boundary conditions predicted by the power PC theory do avoid the ambiguities just noted, as a result of either the dependent measure used or the causal knowledge induced by the participants before the evaluation of the critical candidate cause. In this section, I review several such studies based on three well-known designs borrowed from Pavlovian conditioning: blocking, overexpectation, and conditioned inhibition.

The Boundary Conditions for Evaluating Generative Power as Manifested in Blocking

To avoid the ambiguities found in the previous studies, Fratianne and Cheng (1995, Experiment 3) tested a variant of the blocking design. Unlike previous blocking studies, they (a) specified the type of causal power that participants were to assess, in their case generative power only, and (b) used a response scale that allowed participants to explicitly differentiate among certainty that a candidate cause is noncausal, uncertainty that a candidate is causal, and certainty that a candidate is causal.

Fratianne and Cheng (1995) presented each participant in their experiment with two cover stories, each of which involved assessing the generative nature of three unfamiliar candidate causes. For example, one of the stories concerned helping a botanist determine whether chemicals A, B, and C in fertilizers produce a certain effect (e.g., the growth of a root fungus in plants). Each of the two information patterns summarized in

Figure 2 was presented in the context of each of the cover stories. The information indicated in each cell of the figure, presented to the participants as an item in a randomized list, applies to a group of trials (e.g., plants). (A question mark in a cell in the figure indicates that no information regarding that cell was presented.) Pattern 1 contains the pattern of information in the traditional blocking design, where A is the blocking cue and B is the to-be-blocked cue. (Unlike the traditional blocking design, this experiment presented all information in the same phase.) Participants were asked whether a candidate "causes" a certain effect, and they answered by rating each of the three candidates on a scale that ranged from − 100 (completely confident that the candidate is not a cause) to + 100 (completely confident that the candidate is a cause), with a rating of 0 indicating no confidence at all about the causal status of the candidate.

Part of this experiment replicated other studies but removed the previous interpretative ambiguities. Recall that the power PC interpretation of previous comparisons involving a noncontingent candidate requires the assumption that when participants are uncertain of the causal status of the candidate, they rate it as less causal than a contingent one. The comparison between candidate B in Pattern 1 and candidate Y in Pattern 3 was designed to show that reasoners select focal sets in which alternative causes are controlled without this assumption. The unconditional contrasts for these candidates were equal: As can be seen in Figure 2, information in the left column of each contingency table for Pattern 1 (i.e., $P(e \mid B)$), is equivalent to that for Pattern 3 (i.e., $P(e \mid Y)$); and likewise for information in the right columns (i.e., $P(e \mid \bar{B}) = P(e \mid \bar{Y})$). Candidate B, however, had a contrast of zero conditional on the presence of alternative cause A (i.e., $P(e \mid AB) - P(e \mid A\bar{B}) = 0$; see the top four cells across the two contingency tables in Pattern 1); the corresponding candidate in Pattern 3, Y, had a contrast of one conditional on the absence of alternative cause X (i.e., $P(e \mid \bar{X}Y) - P(e \mid \bar{X}\bar{Y}) = 1$; see the bottom four cells across the two contingency tables in Pattern 3). These were the only contrasts available for candidates B and Y for which alternative plausible causes occur independently of the candidate.[7] Because participants were unambiguously asked to evaluate the generative power of the candidates, Equation 8 unequivocally applies, indicating that the conditional contrast of zero for B does not reveal causal power (because the effect always occurs even in the absence of B). In contrast, the conditional contrast of one for Y does. Thus, whereas the traditional contingency model predicts that participants should be equally confident that B and Y are causal, the power PC theory predicts that they should be uncertain of the causal status of B and, therefore, more confident that Y is causal than they are that B is. Refuting the traditional contingency model, participants clearly and reliably judged Y as causal (with a mean confidence rating of 92) more confidently than they judged B (with a mean rating of − 20). These results using our dependent measure confirm the previous finding that reasoners select focal sets in which alternative causes are controlled.[8]

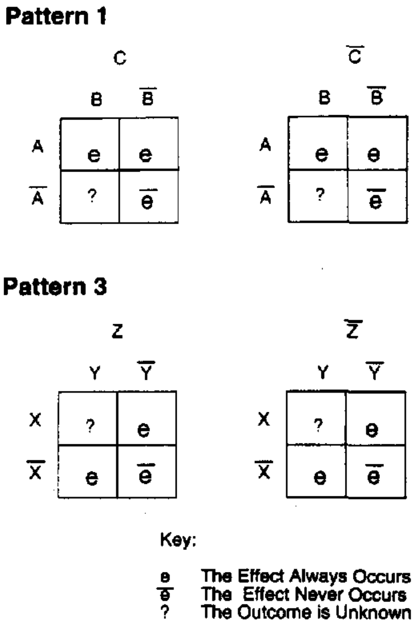


Figure 2. Pattern 1 and Pattern 3 outcome information in Fratianne and Cheng (1995, Experiment 3). The contingency tables indicate whether or not the effect occurs for the various combinations of candidate causes A, B, and C.

Recall that previous studies cannot differentiate between candidate causes with zero contrasts that do or do not reveal causal power according to the power PC theory. To test this novel prediction of the theory, Fratianne and Cheng (1995) created multiple focal sets in which the conditional contrast for a candidate equals 0 but not all of which reveal causal power. Candidate C in Pattern 1 and candidate Z in Pattern 3 both had a contrast of zero conditional on the absence of alternative causes (see the lower right corners of each contingency table in the two patterns). Because the effect did not occur in the absence of C or Z, the contrasts for these candidates in these focal sets should be good estimates of their causal powers according to Equation 8. Now, recall that the only available conditional contrast for candidate B in Pattern 1 was also zero, but this contrast does

[7] As explained later, C in Pattern 1 and Z in Pattern 3 are noncausal according to the power PC theory. They therefore need not be constant in the focal sets for B and Y, respectively. The conditional contrasts for B and Y remain unaltered, however, by including, respectively, C in Pattern 1 and Z in Pattern 3 as alternative causes and conditionalizing on their absence.

[8] Like Spellman (1996a), Fratianne and Cheng (1995) also varied unconditional contrasts while controlling for conditional contrasts. Candidate A in Pattern 1 and candidate X in Pattern 3 had equal contrasts conditional on the absence of alternative plausible causes, the contrast predicted to optimally reveal causal power: They each had a contrast of 1 conditional on the absence of alternative causes (see the right columns of the two contingency tables for each pattern). The unconditional contrast for these candidates, however, differed: This contrast was 1 for A but less than 1 for X. Thus, for this pair of candidates, whereas the traditional contingency model predicts that A should be rated causal with higher confidence, the power PC theory predicts that the two candidates should be rated causal with equal confidence. The mean ratings of these candidates were highly similar, as predicted by their conditional contrasts according to the power PC theory. The mean rating of A (95) was not reliably different from the mean rating of X (93). Differences in unconditional contrast did not matter.

not reveal causal power. The power PC theory therefore predicts that participants should judge C and Z as noncausal more confidently than they judge that B is. This is what Fratianne and Cheng found: Even though all three candidates, B, C, and Z, had conditional contrasts of zero, participants were reliably more confident that C (mean rating = −88) and Z (mean rating = −91) were noncausal than they were that B was (mean rating = −20). This pattern of results contradicts the predictions of the R–W model. One might note that the differences in unconditional contrast among these candidates do predict the pattern of results for this set of candidates. Only the power PC theory, however, successfully predicts the pattern of results for all of the candidates.

In sum, Fratianne and Cheng's (1995) results show that, rather than basing their causal judgments on the entire set of events presented by the experimenter, participants selected focal sets in which alternative causes were controlled. When a contrast in such a focal set revealed causal power, participants confidently judged whether a candidate was causal or noncausal according to its value. When a contrast in such a focal set did not reveal causal power, however, participants were unsure of the causal status of the candidate (as was the doctor in my allergy anecdote). These results, which are based on a dependent measure that removes the ambiguities in previous studies, clearly support the power PC theory and contradict the R–W model. In particular, they support the power PC theory's novel prediction regarding a boundary condition for interpreting positive contrasts.

A Boundary Condition on "Overexpectation" Resulting From Combining Generative Causes

Fratianne and Cheng's (1995) experiment tested the boundary conditions for assessing generative power. Going beyond previous studies, they tested the boundary condition specifying $P(e|\bar{i}) < 1$ by comparing, among focal sets in which alternative causes occur independently of a candidate cause, (a) one in which the effect always occurs even in the absence of the candidate (B in Pattern 1) and (b) one in which the effect never occurs in the absence of the candidate (e.g., Z in Pattern 3). A more stringent test of this boundary condition is to compare the former situation with one in which the effect only sometimes occurs in the absence of the candidate. A particularly interesting design for such a test is based on the overexpectation design (Kremer, 1978; Rescorla & Wagner, 1972) in the Pavlovian conditioning literature. This design involves two learning phases. In Phase 1, stimuli A and B are separately established as excitors ($A+$ and $B+$): Whenever A or B is present, the outcome occurs; on other trials, no stimulus occurs, and the outcome also does not occur. In Phase 2, the combination of A and B is shown to also lead to the effect (AB +) to the same degree as did each of the individual stimuli in Phase 1: Whenever the combination of A and B occurs, the outcome occurs; on other trials, when no stimulus occurs, the outcome also does not occur. Tests subsequent to Phase 2 typically reveal that the excitatory power of both A and B has decreased relative to Phase 1. The R–W model predicts this result because the sum of the associative strengths of A and B after Phase 1 "overpredicts" the outcome associated with the AB compound in Phase 2, leading to reduction in both weights.

Cheng, Park, Yarlas, and Holyoak (1996) explained that, from the perspective of the power PC theory, this result depends on the fact that the laboratory-animal studies have used events (e.g., shock administered at various times) that are naturally interpreted as occurring with certain rates in continuous time rather than with certain probabilities (Gallistel, 1990). Rates have no a priori upper bound (except that imposed by technology and perceptual systems), in contrast to probabilities, which have an upper bound of 1. For example, an experimenter might define a trial to be of a certain length in time (e.g., 2 min), with a shock occurring once every trial as indicating that it occurs with a probability of 1. However, it is physically possible for shocks to occur at rates higher than this artificial upper bound and be perceived as such up to the limit imposed by the participants' perceptual system. (Like probabilities, however, rates do have a lower bound, it is physically impossible for effects to occur at a rate slower than 0 per unit time interval.) For generative causes, therefore, rates do not have an analogue of the boundary condition specifying $P(e|\bar{i}) < 1$, a condition that applies to probabilities rather than rates. The failure of the AB compound in Phase 2 to increase the rate beyond that associated with each cue alone is thus normatively interpreted in terms of rates as evidence that neither cue is as potent as it had appeared in Phase 1.

Discrete effects (e.g., being pregnant) that occur in discrete trials (e.g., a woman), however, are naturally coded in terms of probabilities rather than rates. The boundary condition for generative causes predicted by the power PC theory should therefore apply. As I explain later, this theory predicts that whether overexpectation occurs should depend on whether the candidates are paired with the effect all of the time (the "ceiling" case) or only sometimes (the "nonceiling" case). In contrast, the R–W model, as a result of its additivity, does not distinguish between these two cases and predicts overexpectation in both.

Estimating generative causal power in a ceiling and a nonceiling situation. First, consider the ceiling situation. Recall that in Phase 1, when either candidate A or B is present, the effect always occurs; otherwise, the effect does not occur. This design implies that, for each candidate cause, there is a focal set in which causes alternative to the candidate occur independently of it. With respect to the assessment of candidate A, this focal set consists of trials on which B is absent. In this set, the only alternative cause is the "context," which is constant and therefore independent of A. Therefore, either Equation 8 or 14 applies. Because the contrast for A in this set is positive, the choice is Equation 8. For candidate A, $P(e|\bar{i})$ in Equation 8 equals 0, because the effect does not occur when the context alone is present. Under these conditions, this equation indicates that this contrast, which equals 1, provides a good estimate of the causal power of A. The same reasoning applies to candidate B. Therefore, if reasoners select focal sets that reveal causal power, they should rate both A and B as highly causal in Phase 1.

Accumulating information across the two phases allows the selection of a second type of focal set in which causes alternative to a candidate are constant. With respect to the assessment of candidate A, this additional focal set consists of trials on which

B and the context are always present. In this focal set, the contrast for A, $P(e|ABC) - P(e|\bar{A}BC)$, where C is the context, is 0: Both conditional probabilities equal 1. (Recall that in Phase 1 in the ceiling situation, when A or B is present, the effect always occurs; likewise, in Phase 2, when the AB compound is present, the effect always occurs.) According to the power PC theory, however, this contrast is uninterpretable as an estimate of causal power; because the effect always occurs in the absence of A, $P(e|\bar{i})$ in Equation 8 equals 1. Equation 8, the equation for evaluating generative causal power, is relevant because A has been inferred to be generative in Phase 1. Thus, in this case, prior causal knowledge specifies which equation is relevant. The same reasoning applies to candidate B, for which the analogous conditional contrast is likewise 0 and likewise uninterpretable. Therefore, although the contrast for each candidate changes from 1 in the focal set used in Phase 1 to 0 in the focal set available in Phase 2, there is no conflict in the interpretation according to causal power. Accordingly, the power PC theory predicts that reasoners will not change their causal assessment of either A or B during Phase 2. That is, contrary to the prediction of the R–W model, overexpectation will not occur in the ceiling situation.

To serve as a baseline for comparison, consider a control group that uses an identical design and procedure as the experimental group, except that one of the two candidate causes, A, is not paired with the effect in Phase 1. According to the power PC theory, A should be noncausal: $P(e|A\bar{B}) - P(e|\bar{A}\bar{B}) = 0$. However, B should have the same causal power in the experimental and control groups because its contrast conditional on the absence of A has the same value (namely, 1) and the same base rate of the effect (namely, 0) in both groups; therefore, these groups should not differ in their assessment of B in Phase 1. More important with respect to the prediction of overexpectation, because A is noncausal in the control group, the relevant contrast for B is the unconditional one, which does not change across phases; in both phases, $P(e|B) = 1$ and $P(e|\bar{B}) = 0$. Therefore, this model predicts that the estimated causal strength of B should not change across phases (i.e., there should be no overexpectation) in the control group, and thus there should be no difference between the two groups.

Now consider a nonceiling situation. The only difference in design between this situation and the previous one is that in those stimulus contexts in which the effect occurs (for both the experimental and control groups), rather than always occurring, it occurs with a constant probability that is clearly greater than 0 and less than 1. It follows that, for the experimental group in this situation, the contrast for either A or B in Phase 1 conditional on the absence of the other is clearly between 0 and 1. These contrasts should be good estimates of generative causal power, because the only alternative cause in these focal sets is the context, which does not produce the effect. The causal power of either A or B should therefore be clearly less than 1. Now, in Phase 2 of the nonceiling situation, as in the ceiling situation, the contrast for either candidate conditional on the presence of the other is 0, because the effect occurs with the same frequency in the presence of the AB compound as when either candidate is present in the earlier phase. Contrary to those in the ceiling situation, however, these contrasts should provide an estimate of causal power: $P(e|\bar{i})$ in Equation 8 is clearly less than 1 in these focal sets (because the probability of the effect produced by alternative cause A or B is clearly less than 1). Therefore, a conflict appears between the causal powers inferred in the two phases.

There are multiple ways of resolving this conflict, not all of which would result in overexpectation under a power analysis. Consider a reasoner who assumes that the candidate causes have stable causal powers and who trusts the observations in Phase 2 but doubts those in Phase 1. This reasoner could resolve the conflict by hypothesizing that the proportion of cases in Phase 1 for which A or B produced the effect is lower than originally estimated. Such a modification would result in lower estimates of causal power for A or B, which would be consistent across the two phases. Accordingly, the power PC theory predicts that overexpectation will occur in the nonceiling situation if the experimental procedure encourages participants to trust the observations in Phase 2 but doubt those in Phase 1; that is, the causal ratings for B in Phase 2 should be lower in the experimental group than in the control group.

In sum, the power PC theory predicts a possible reduction in the perceived causal power of A and B during Phase 2 of the overexpectation design using discrete trials when the effect occurs with a nonceiling probability in the presence of A, B, or their combination but no such reduction when the effect occurs with a ceiling probability of 1 in these stimulus contexts.

The overexpectation design not only tests the boundary condition specifying $P(e|\bar{i}) < 1$ but also tests the selection of focal sets. Note that the unconditional contrast for B in Phase 1 is lower in the experimental group than in the control group in both the ceiling and nonceiling situations: Because A (which appears in the absence of B) is paired with the effect only in the experimental group, $P(e|\bar{B})$ is higher in this group. In addition, the unconditional contrast does not change for either A or B across phases in either the ceiling or the nonceiling situation. If participants do not select focal sets in which alternative causes occur independently of the candidate cause, B in Phase 1 should be rated less causal in the experimental group than in the control group in both the ceiling and the nonceiling situations. Moreover, causal judgments regarding A and B should not change across phases, for the experimental group in the nonceiling situation as well as for other groups. These predictions based on unconditional contrasts contradict those derived from the power PC theory.

Power analysis of the contrasts computed by the R–W model in the overexpectation design. The R–W model's prediction of overexpectation rests on a comparison between the two phases of the overexpectation design. Recall that Phase 1 has the following design: $C-$, $AC+$, and $BC+$ (where C is the context cue). This design is nested because C, the intersection of the partially overlapping combinations AC and BC, occurs by itself as a separate combination. According to my derivation, this model computes the contrasts for A or B in this design conditional on the presence of the context alone. Because the context does not produce the effect (i.e., $P(e|\bar{i})$ equals 0), these contrasts closely estimate causal power according to Equation 8. The prediction of the R–W model therefore coincides exactly with that of the power PC theory in this case.

In Phase 2, the design is $C-$ and $ABC+$. Because ABC is a superset of C, this design is also nested. For this design, the model computes the contrast for *AB* as a composite conditional on the presence of the context alone. Because the effect occurs with equal probability in the presence of noncontext stimuli in this design, the contrast for this composite has the same value as that for the individual stimuli in Phase 1. The additivity assumption of the R–W model therefore yields the prediction that the strengths of the individual stimuli in Phase 2 will asymptotically be reduced to half their previous strengths. For the nonceiling experiment, this direction of change is consistent with that predicted by a hypothesized revision of observations in Phase 1 to obtain a consistent estimate of causal power across phases. But recall that, with regard to the ceiling experiment, the only contrast that reveals the causal power of *A* or *B* is that obtained in Phase 1. Because the prediction of the R–W model for Phase 2 does not correspond to this contrast (instead, it predicts a value half the magnitude of this contrast), a power analysis predicts that the model should fail.

Experimental test of overexpectation using discrete trials. To test the predicted selection of focal sets and interpretability of zero contrasts derived from the power PC theory, Park and Cheng (1995) conducted two overexpectation experiments using discrete trials with humans performing a causal induction task. These experiments were identical in design except for the theoretically crucial distinction between whether the critical stimulus contexts produce the effect sometimes (nonceiling experiment) or always (ceiling experiment). College students were presented with a cover story in which they were asked to infer how likely it is that various newly discovered (fictitious) proteins called "endomins," which were said to sometimes be produced by the body, caused hypertension in people who have those endomins. Trials therefore consisted of people, who are discrete entities.

The two experiments respectively used the designs just described for the ceiling and nonceiling situations: For the ceiling situation, the probability of hypertension given any pattern of endomins was 1 whenever it was positive in the designs described; for the nonceiling situation, these probabilities were .75. On each learning trial, participants were given a "hospital record" listing information about the presence or absence of three endomins and of hypertension in a particular patient.[9] None of the participants had been exposed to probability theory.

As a means of measuring the participants' causal judgments, they were given a response sheet that listed patterns of endomins (the three individual endomins, the combination of *A* and *B*, and no endomins). For each pattern, participants were asked, "Out of 100 patients with this pattern of endomins, how many do you think have hypertension?"

Recall that for the discrete-trial version of the overexpectation design, the power PC theory predicts overexpectation for reasoners in a nonceiling situation who trust the observations presented in Phase 2 but doubt those in Phase 1. To create an experimental procedure allowing the power PC theory to make different predictions for a ceiling and a nonceiling situation, the instructions encouraged all participants to doubt the observations in Phase 1 if they perceived conflicting information in the two phases. At the beginning of Phase 2, every participant was told that there may or may not be some inaccurate diagnoses of patients whose records they have just seen but that the diagnoses of the patients whose records they were going to see were certainly accurate.

The results of both experiments were in accord with the predictions of the power PC theory. As predicted, on one hand, in the experimental group of the nonceiling experiment, the mean number of patients (out of 100) who were estimated to have hypertension for the critical stimulus *B* decreased reliably (by 10) across the two phases. This reduction was reliably greater than the corresponding decline (of 2) observed in the control group, a decline that was not reliable. On the other hand, in the ceiling experiment, the estimates for *B* were virtually unchanged across phases (ranging from 99 to 100) in both the experimental and control groups. The number of participants who did or did not show a reduction in their estimate for *B* in Phase 2 indicated the same pattern of results. In the nonceiling experiment, a reliably larger proportion of participants in the experimental group than in the control group showed a decline in their estimate for *B* from Phase 1 to Phase 2. In contrast, in the ceiling experiment, there was no difference between groups; none of the participants in either group showed such a decline.

In sum, as predicted by the power PC theory, but no other theory, stimulus *B* lost perceived causal power in Phase 2 for the experimental group (but not the control group) of the nonceiling experiment. In contrast, the perceived causal power of this stimulus remained unchanged in both groups of the ceiling experiment. Also as predicted by the power PC theory, the experimental and control groups rated *B* similarly in Phase 1, in both the nonceiling experiment (78 and 74 for the two groups, respectively) and the ceiling experiment (99 for both groups), despite the lower unconditional contrast for *B* in the experimental groups.

Implications. The results obtained by Park and Cheng (1995) provide strong support for the interpretation of contrasts as estimates of causal power. The power PC theory accurately predicts a boundary condition on the phenomenon of overexpectation for human causal induction with events that are naturally encoded in terms of probabilities. Depending on whether each of two individual cues sometimes (nonceiling experiment) or always (ceiling experiment) produced the effect, the estimated causal power of each individual cue was, respectively, (a) reduced (as a result of overexpectation) or (b) not reduced when the two cues were presented in combination with the same probability of the effect as had been observed for each cue alone. Recall that Pavlovian conditioning studies using the same design as the ceiling experiment did obtain overexpectation. The critical difference is that these conditioning studies presented events that occurred with certain rates rather than probabilities, and an overexpectation design implemented with such events should

[9] Our only modification to the standard overexpectation design was the addition of a third stimulus. This stimulus was always presented in isolation; therefore, its addition should not affect the strengths of other stimuli according to both the R-W model and the power PC theory (see Appendix A). This stimulus was always followed by the effect with a probability of .92 in both phases for both groups in the nonceiling situation and with a probability of 1 in both phases for both groups in the ceiling situation. It was added to provide a comparison for confirming that a probability of .75 was perceptibly below a near ceiling level (it was).

yield overexpectation according to an analysis of causal power. Without an interpretation in terms of causal power, no model of covariation (including the R–W model) is able to account for the observed differences in overexpectation (a) between the ceiling and nonceiling experiments and (b) between the ceiling experiment and the conditioning experiments.

Moreover, Park and Cheng's (1995) experiments provide further support for the selection of focal sets that optimally reveal causal power. The overexpectation observed in their nonceiling experiment cannot be explained in terms of unconditional contrasts, which did not change across phases. Neither can such contrasts explain why participants in the experimental and control groups rated B similarly in Phase 1 for both experiments, despite the difference in unconditional contrast for this candidate between groups.

A Boundary Condition on the Interpretation of Inhibitory Causes

Conditioned inhibition. Whereas the overexpectation design permits a test of the boundary condition particular to the assessment of generative power predicted by the power PC theory, a causal analogue of another conditioning design permits a test of the corresponding boundary condition for assessing preventive power predicted by this theory: the extinction of *conditioned inhibition* (or, in causal terms, the reduction in perceived power of a preventive cause). The initial acquisition of conditioned inhibition was first described by Pavlov (1927). In this procedure, some outcome occurs in the presence of a stimulus A ($A+$) but not in its absence; neither does this outcome occur when A is paired with a second stimulus X ($AX-$). If we let C represent the context, then the design for conditioned inhibition is $C-$, $CA+$, and $CAX-$. Exposure to these events leads to A being perceived as predicting the outcome and X as inhibiting it. This perception of inhibition can be behaviorally tested via a transfer task known as summation. Pavlov (1927) showed that when X is later paired with some novel excitatory stimulus B, the response that had been previously evoked by B is attenuated. The summation of B and X indicates that the learner possesses a generalized inhibitory representation of the X stimulus independent of the stimulus with which it was originally paired.

The power PC theory predicts that reasoners will judge X to be inhibitory in this design, consistent with Pavlov's (1927) and Rescorla's (1969) findings using animals and Williams's (1995) and Williams and Docking's (1995) findings using humans with inference tasks. The only contrast for X when alternative causes are controlled (i.e., $P(e|CAX) - P(e|CA\bar{X})$ is negative. Equation 14—the equation for evaluating inhibitory power—therefore applies, yielding the prediction that X would become an inhibitor. This contrast is also the one computed by the R–W model for X. Because every stimulus combination in this design except the one with a single stimulus can be characterized as a superset of all combinations with fewer stimuli, the design is nested, and the model computes the contrast for X conditional on the cues in the next smaller combination, CA. The R–W model therefore makes the same prediction as the power PC theory.

Extinction of conditioned inhibition. The extinction of a conditioned inhibiting stimulus (such as X described earlier) occurs when new information leads to X no longer being perceived as preventive. Under a "direct" procedure, the conditioned inhibiting stimulus X is subsequently presented alone with no outcome ($X-$). Letting C represent the context as before, the design is $C-$ and $CX-$. The intervening experience with X in the absence of excitatory cause A yields the contrast $P(e|C\bar{A}X) - P(C\bar{A}\bar{X}) = 0$. Because the design is nested in this phase, this contrast is the one computed by the R–W model for X. This model therefore predicts that the inhibitory power of X will become extinguished. According to the power PC theory, however, this contrast is uninterpretable as an estimate of the inhibitory power of X: Prior causal knowledge about X indicates that Equation 14 is relevant, and, for this contrast, $P(e|\bar{i})$ in this equation equals 0. The new information therefore does not conflict with the estimate for X obtained in the earlier phase. Accordingly, this intervening experience will not alter the previous estimate, and the direct procedure will not lead to the extinction of conditioned inhibition.[10]

Experiments using this design with both humans and laboratory animals have supported this prediction of the power PC theory, contradicting that of the R–W model. Zimmer-Hart and Rescorla (1974) found that when a conditioned inhibitory stimulus (a light flash) was presented alone with no outcome, it retained its inhibitory strength in later summation trials when paired with a novel excitatory stimulus (a tone). Yarlas et al. (1995) replicated this pattern of results on a summation test using humans with a causal inference task.

Note that the preceding prediction of the power PC theory is the preventive analog of the theory's prediction of lack of overexpectation when generative causes that produce the effect at a ceiling level are combined. Recall that in Phase 2 of the overexpectation design, information regarding the critical candidate in the presence of an alternative generative cause yields a contrast that is uninterpretable as an estimate of the candidate's generative power according to Equation 8, because the effect always occurs in the presence of the alternative cause. The power theory therefore predicts (correctly) that this contrast will be ignored in favor of an available contrast that is interpretable. Analogously, in the extinction phase of the conditioned inhibition design, experience with X in the absence of generative cause A yields a contrast that is uninterpretable as an estimate of the inhibitory power of X according to Equation 14, because the effect never occurs in the absence of A. Reasoners are therefore predicted to ignore this contrast in favor of an interpretable one obtained earlier.

Counterintuitively, the power PC theory predicts that the inhibitory power of X will be extinguished in an "indirect" procedure, in which the previously generative candidate A, which had been inhibited by X, is at a later time no longer paired with the presence of the outcome (i.e., $A-$). According to the power PC theory, the relevant conditional contrast, $P(e|CAX) - P(e|CA\bar{X})$, which had been negative when A was generative, will be reduced given the subsequent events (the value of the

[10] The unconditional contrast for X is also unaffected by the direct extinction procedure. However, it is the relevant conditional contrast that is crucial, as was observed in Fratianne and Cheng (1995), Park and Cheng (1995), and other studies reviewed in the overshadowing and blocking sections.

first term remains at 0, whereas the value of the second shifts from 1 toward 0). Several studies of animal conditioning, extinguishing either the conditioning context or some previously generative candidate (Best, Dunn, Batson, Meachum, & Nash, 1985; Hallam et al., 1990; Kaplan & Hearst, 1985; Kasprow, Schachtman, & Miller, 1987; Lysle & Fowler, 1985; Miller & Schachtman, 1985), have yielded results consistent with the selection of this optimal contrast: Conditioned inhibition was extinguished via the indirect "retrospective" procedure. Williams and Docking (1995) and Yarlas et al. (1995) replicated this finding with humans using causal inference tasks.

For this retrospective procedure, as for other retrospective procedures, the R–W model fails. It does not revise the strength of inhibitory stimulus *X* in the extinction phase of this procedure because the stimulus is absent. This stimulus therefore retains the value of its contrast from the acquisition phase rather than changing to its updated value in the extinction phase, the value that is informative according to the power PC theory.

Implications. With respect to the extinction of conditioned inhibition, just as with overexpectation, it thus appears that the power PC theory provides an accurate model of causal inference as well as Pavlovian conditioning. Both the failure of the direct procedure to "extinguish" a preventive cause in Yarlas et al. (1995) and the success of the indirect extinction procedure in Williams and Docking (1995) and Yarlas et al. (1995) strongly support the predictions of the power PC theory. Confirming the boundary condition for evaluating preventive causal power predicted by the power PC theory, the direct procedure failed to extinguish such a cause. As in the case of the hypothetical test of the headache-relieving drug, the zero contrast for a candidate is uninterpretable when no causes are producing the effect. Preventive causes are instead extinguished indirectly by extinguishing the generative cause that the preventive cause previously inhibited, confirming the optimal contrast for assessing the preventive power of a candidate cause predicted by this theory.

At the same time, these results are diametrically opposite to those predicted by the R–W model. The failure of the direct procedure illustrates that treating a preventive cause as the mirror image of a generative cause, as this model does, is problematic. Whereas the strength of a generative cause can be reduced by presenting it without the effect (Williams, 1995; Williams & Docking, 1995; Yarlas et al., 1995), the strength of a preventive cause cannot be reduced using the same procedure, as predicted by the different optimal focal sets for evaluating generative and preventive power. The success of the indirect extinction procedure illustrates that the R–W model's assumption that the strength of a cue is not updated when it is absent is problematic when the updated value does reveal causal power.

Summary of Studies Involving Multiple Varying Candidate Causes

Studies testing the traditional contingency model using the blocking and induced overshadowing designs uniformly support the power PC theory's prediction regarding the selection of focal sets in which alternative causes occur independently of a candidate cause: Causal judgments were observed to depend on contrast within such focal sets—the focal sets that potentially reveal causal power—rather than within the universal set that consisted of all information provided by the experimenter. These studies, however, do not allow an evaluation of this theory's predictions regarding the different boundary conditions for interpreting generative and preventive powers within focal sets in which alternative causes are controlled, predictions that discriminate between the power PC theory and the R–W model. Studies that do allow such an evaluation support the power PC theory. On one hand, Fratianne and Cheng (1995) and Park and Cheng (1995), respectively using the blocking design and the overexpectation design, tested the predictions of this theory regarding nonnegative contrasts as estimates of generative causal power. Participants based their causal judgments on contrast, including a zero contrast, in focal sets in which alternative causes were controlled when these causes were sometimes (Park & Cheng, 1995) or never (Fratianne & Cheng, 1995) producing the effect. In contrast, they either ignored a zero contrast in such a focal set when an alternative cause was always producing the effect (Park & Cheng, 1995) or were uncertain of its causal implications if such a contrast was the only one available (Fratianne & Cheng, 1995). On the other hand, Williams and Docking (1995) and Yarlas et al. (1995), using the conditioned inhibition design, tested the predictions of the power PC theory regarding nonpositive contrasts as estimates of preventive causal power. Participants based their causal judgments of preventive causal power on contrast, including a zero contrast, in focal sets in which alternative causes were controlled when the effect always or at least sometimes occurred in the absence of the candidate, whereas they ignored a zero contrast in such a focal set when the effect never occurred in the absence of the candidate.

My review shows that when the R–W model computes a conditional contrast that reveals causal power (and hence makes the same prediction as the power PC theory), it is supported empirically, as in studies of conditioned inhibition, of induced overshadowing, and of overexpectation using continuous trials; a study of overexpectation using discrete trials when $P(e|i) < 1$ for both candidates in this design (i.e., the effect only sometimes occurred in the presence of each of the two candidate causes); and studies of blocking in which the dependent measure did not discriminate between certainty of noncausality and uncertainty of causal status. Otherwise, this model is not supported, as in studies of the extinction of conditioned inhibition, a study of overexpectation using discrete trials when $P(e|i) = 1$ for both candidates in this design, and a blocking study that does allow the discrimination between certainty of noncausality and uncertainty of causal status.

Causal Estimates as a Joint Function of Contingency and the Base Rate of the Effect: Tests of the Traditional Contingency Model Involving a Single Varying Candidate Cause

I now turn to the evaluation of the power PC theory's predictions for studies involving a single varying candidate cause (other causes are held constant). Such studies show that the magnitude of observed causal ratings is a function of (a) contingency, as predicted by the traditional contingency model, the R–W model, and the power PC theory, and (b) the base rate of the effect, as predicted by the power PC theory alone.

Many studies have tested the traditional contingency model using designs involving a single varying candidate cause i by orthogonally varying the probability of an effect given the presence of the candidate, $P(e|i)$, and this probability given its absence, $P(e|\bar{i})$. These studies have uniformly reported that the magnitude of observed causal ratings was a function of the contingency for the candidate cause C conditional on the constant presence of a context cue X (e.g., Anderson & Sheu, 1995; Baker et al., 1989; Shanks, 1985a, 1987, 1995; Wasserman et al., 1983, 1993). In the most comprehensive study to date involving a single varying candidate, in which Wasserman et al. (1993) varied five values of $P(e|i)$ and $P(e|\bar{i})$ orthogonally, the correlations between mean observed causal ratings and contingency were .97 and .98 in two experiments. These results strongly support the traditional contingency model.

When the test situation involves a single varying candidate cause, this model's predictions coincide entirely with the asymptotic predictions of the R–W model if its learning rate parameters are assumed to be constant across trials on which the effect does or does not occur (Appendix A; Chapman & Robbins, 1990). Because the context (X), the only alternative cause, is constantly present, the unconditional contrast for C ($P(e|C) - P(e|\bar{C})$ is equivalent to its contrast conditional on the presence of X (i.e., $P(e|XC) - P(e|X\bar{C})$. In other words, the universal focal set in these studies is a focal set that reveals causal power. The two cue combinations in this design are XC and X, which are nested. From my derivation, it follows that the strength of C according to the R–W model is equal to the contrast for C conditional on the presence of X (Equation A23 in Appendix A). As I show in a later section, the power PC theory predicts that as long as the boundary conditions indicated by Equations 8 and 14 are satisfied, when contingency is manipulated by changing either the frequency of the effect in the presence of the target cue or that in its absence, as in the studies just cited, the change in estimated causal power of the target cue is always in the same direction as the change in its contrast (i.e., $P(e|XC) - P(E|X\bar{C})$ and proportional to it. The predictions of the R–W model, which are identical to those of the traditional contingency model, therefore coincide ordinally with those of the power PC theory. In sum, the results supporting the traditional contingency model equally support the power PC theory and the R–W model: They converge in predicting that causal judgments are a function of contingency.

Despite the impressive accordance between contingency and the observed values, researchers have reported a "bias" due to the probability of the effect ($P(e)$) for candidate causes with the same value of ΔP, both for candidates with a contingency of zero (Allan & Jenkins, 1980, 1983; Alloy & Abramson, 1979; Baker et al., 1989; Chatlosh et al., 1985; Dickinson et al., 1984; Shanks, 1985a, 1987) and for candidates with a nonzero contingency (Allan & Jenkins, 1983; Wasserman et al., 1983, 1993). I consider judgments regarding noncontingent and contingent candidates separately, because they have different interpretations according to the power PC theory.

Bias in Judging Noncontingent Candidates

For noncontingent candidates, participants in several studies reliably judged candidates as more positive when the probability of the effect was high than when it was low (e.g., Baker et al., 1989; Chatlosh et al., 1985; Dickinson et al., 1984). For example, Baker et al. (1989, Experiment 3) reported that when $P(e)$ was .75, the mean rating given to a noncontingent candidate was positive, but when $P(e)$ was .25, this mean rating was negative, and the two mean ratings differed reliably. (Note that for noncontingent candidates, $P(e) = P(e|i) = P(e|\bar{i})$.) All studies reporting such a bias used a within-subject design in which the same group of participants judged candidates with positive, negative, and zero contingencies. For each candidate, participants were not restricted to assessing either its generative or its inhibitory nature.

The reported influence of $P(e)$ on the judgment of noncontingency candidates might appear to contradict the power PC theory (it contradicts the R–W model and the traditional contingency model). I show later, however, that this "bias" is consistent with the power PC theory if (a) there are variations in the objective value of ΔP from participant to participant (with a mean of 0 across participants) or (b) participants are likely to misperceive an objective ΔP of 0. One or both of these conditions obtained in all of the studies in which this bias was reported. (The zero contingencies in these studies were unlikely to be perceived as undefined, because $P(e|\bar{i})$ did not approximate 0 or 1.) Recall that a nonzero ΔP (when its boundary conditions are satisfied) provides an increasingly conservative estimate of causal power in opposite directions as a function of $P(e|\bar{i})$ depending on whether one is evaluating generative or preventive power. As a result, participants might have selectively exaggerated any objective or subjective deviation of the value of ΔP from zero depending on the direction of the deviation, as explained below.

Suppose that a participant perceives a purportedly noncontingent candidate to have a small positive ΔP. When $P(e)$ is high, and hence $P(e|\bar{i})$ is high, ΔP underestimates the generative power of the candidate (Equation 8); consequently, the participant should compensate for the underestimation by giving an estimate that is higher than the perceived positive value of ΔP. When $P(e|\bar{i})$ is low, however, ΔP closely estimates the generative power of the candidate. Therefore, the participant in this situation should give the perceived value of ΔP directly as an estimate.

Conversely, suppose that a participant perceives a purportedly noncontingent candidate to have a small negative ΔP. When $P(e)$ is high, and hence $P(e|\bar{i})$ is high, ΔP closely estimates the preventive power of the candidate (Equation 14); consequently, the participant should give the perceived value of ΔP directly as an indication of the preventive power of the candidate. When $P(e)$ is low, however, ΔP underestimates the preventive power of the candidate. The participant in this situation might then compensate by giving an estimate that is more negative than the perceived negative value of ΔP (to indicate greater preventive power on the given rating scales).

Now, if some participants perceive the ΔP of the candidate to be positive and some perceive it to be negative, and the sign of the perception is independent of whether $P(e)$ is high or low, then across the entire group of participants a higher $P(e)$ would give rise to a higher mean estimate of causal strength: Whereas causal estimates measured when $P(e)$ is high would include exaggerated positive estimates (and unexaggerated negative and zero estimates), those measured when $P(e)$ is low would include exaggerated negative estimates (and unexaggerated positive and zero estimates). When objective or subjective deviations of the value of ΔP from 0 are probable, such as when the number of trials at the time of the assessment is small, this bias is more likely to operate. Moreover, the fact that the same participants judged both the generative and the preventive nature of the candidates would amplify the bias by pulling estimates in opposite directions when $P(e)$ is high (estimates of generative power are pulled up) and when it is low (estimates of preventive power are pulled down).

Shanks and Lopez (cited in Shanks, 1995, and Shanks et al., 1996) were successful in avoiding this bias by using a relatively large number of trials per condition (40) and a between-subjects design in which each participant judged the contingency of only one candidate. The between-subjects design, in addition to the large number of trials, might have produced more accurate judgments: It avoided proactive interference from preceding contingencies, and participants might have paid more attention on the whole because they had less to do. In two of the conditions, the contrast for the candidate cause was zero: In one condition, $P(e)$ was .25; in the other condition, $P(e)$ was .75. According to both Equations 8 and 14, a contrast of zero for the candidate cause in such situations does indicate a lack of power of the candidate to produce the effect. In accord with this prediction, the observed mean causal ratings indeed converged on zero as the trials increased both when $P(e)$ was high and when it was low.

In sum, although causal judgments on candidate causes with zero contingencies have been reported to increase as a function of the base rate of the effect, such an influence need not refute normative accounts. Rather, it follows from the power PC theory when there is reason to believe that participants might have perceived a purportedly noncontingent candidate to have a nonzero contingency. When some likely factors that contribute to this perception were removed, noncontingent candidate causes were indeed judged to be noncausal. That is, for zero as well as nonzero contingencies, asymptotic studies that varied one candidate cause strongly support the prediction made by the traditional contingency model, the R–W model, and the power PC theory: Causal judgments are a function of ΔP.

Predicted Influence of the Base Rate of the Effect on the Magnitude of Estimated Generative and Preventive Causal Powers

Several articles have reported that candidate causes with the same nonzero ΔP were judged differently depending on $P(e)$ (Allan & Jenkins, 1980, 1983; Wasserman et al., 1983, 1993). Because results were not always presented in such a way to allow inferential statistics for testing the influence of $P(e)$, these findings—although intriguing and apparently systematic—are sometimes only suggestive. These deviations from objective ΔP are biases with respect to the traditional contingency model. As I show later, these deviations concerning positive contingencies in studies using discrete trials are also biases with respect to the R–W model under the typical assumption that β, the parameter in the model associated with the outcome, is larger when the outcome is present than when it is absent. In contrast, these deviations are normative according to the power PC theory.

Recall that the power PC theory predicts that causal judgment is a joint function of contingency and the base rate of the effect: For a positive ΔP_i of the same magnitude, as $P(e|\bar{i})$ increases, higher strengths will be inferred for candidate cause i; for a negative ΔP_i of the same magnitude, as $P(e|\bar{i})$ increases, lower strengths should be inferred for i. Wasserman et al. (1993) conducted the most comprehensive study to date relevant to these predictions. On one hand, because Wasserman et al.'s experiments used continuous trials, which are more appropriately represented in terms of rates rather than probabilities, the power PC theory does not directly apply. On the other hand, as I argue later, for the evaluation of preventive power, it does not matter whether events are represented in terms of probabilities or rates: They have analogous interpretations in terms of causal power, and the same prediction regarding the influence of the base rate of the effect applies.

Participants in Wasserman et al.'s (1993) experiments were asked to judge whether tapping a key had any effect on the occurrence of a white light. The light occurred with various probabilities at the end of 1-s sampling intervals. Because seconds on a time scale are not discrete entities, the occurrence of the light is more appropriately represented in terms of rates rather than probabilities. As I mentioned, unlike probabilities, rates do not have an upper bound. For example, contrary to the maximum "probability" of one light flash every second set by Wasserman et al., the light could have flashed at a higher rate. Rates do have a lower bound, however; events cannot occur at a rate slower than not occurring at all. Wasserman et al.'s light, for example, could not have flashed at a rate any slower than 0 times every second.

A power analysis can be applied to events that occur with certain rates, just as I did for probabilistic events. Such an analysis would show consequences of the lower bound of rates for preventive causes but no parallel consequences of an upper bound for generative causes (unless one comes close to the limits of perception). Let me illustrate this with a concrete example. First, consider evaluating the preventive power of a candidate cause of effects that occur with rates. Suppose there is an observable change of − .50 flash per second in rate when a candidate occurs in comparison with when it does not, assuming that alternative causes occur independently of the candidate. When the context (i.e., alternative cause a in my analysis) is producing the flash at the rate of .75 per second, a candidate cause would need to have a power to reduce the rate by two thirds to yield the observable change of − .50 flash per second. By comparison, when the context is producing the flash at the rate of 1.0 per second, the candidate cause would need to have a power to reduce the rate by only a half to obtain the same observable change. Thus, analogous to negative contrasts in the case of effects that occur with probabilities, for the same observable reduction in rate due to a candidate cause, as the base rate of the effect increases, a lower preventive power would be inferred for the candidate.

The pattern of results reported by Wasserman et al. (1993) supports this prediction regarding the evaluation of preventive power. Judged contingencies[11] were systematically less negative

[11] I use the term *contingency* to be consistent with Wasserman et al.'s (1993) report, in the sense of the analog of contingency (which is

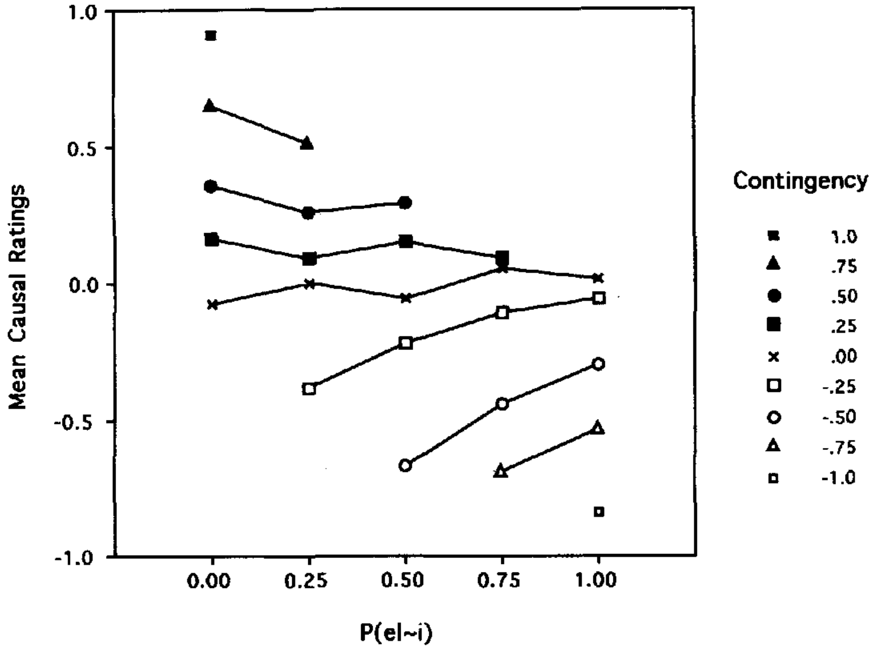


Figure 3. Mean scaled causal rating scores in Experiment 1 of Wasserman et al. (1993) as functions of ΔP and $P(e|\bar{i})$. (Lines join scores with the same contingency.) Adapted from "Rating Causal Relations: The Role of Probability in Judgments of Response-Outcome Contingency," by E. A. Wąsserman, S. M. Elek, D. L. Chatlosh, and A. G. Baker, 1993, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* p. 178. Copyright 1993 by the American Psychological Association. Adapted with permission of the author.

(i.e., judged to have less preventive causal power) for the same objective negative contingency as the rate of the effect in the absence of tapping increased (see the right panels of Figures 1 and 3 in Wasserman et al., 1993; the right panel of their Figure 1 is adapted as Figure 3 here $P(e|\bar{i})$ is denoted as $P(e|{\sim}i)$ in the figure).

In a similar study manipulating $P(e|i)$ and $P(e|\bar{i})$ using fewer values, Wasserman et al. (1983) reported the change in the observed mean causal judgments for candidates with negative contingencies as $P(e|\bar{i})$ increased: For the comparison between pairs of candidate causes with the same negative contingencies in each of the seven experimental groups, the candidate with a higher $P(e|\bar{i})$ was judged less negative. The results reported do not allow paired-comparison t tests, but estimating the pooled variance by assuming that the means came from independent samples, a procedure that should provide conservative estimates of reliability given the within-subject comparisons, the reductions were highly reliable for four of the comparisons, $t(35) = 2.89$ and $t(35) = 2.06$ for the two groups in Experiment 1, $t(17) = 2.95$ for the first group in Experiment 2, and $t(17) = 3.40$ for the second group in Experiment 3 ($p < .05$ for each comparison).

Wasserman et al. (1993) interpreted their results as supporting the R–W model. (This is one of the phenomena I mentioned earlier in which the R–W model requires varying values of β.) Adopting a higher value of β in this model when the outcome is present than when it is absent, in which case the R–W model is equivalent to a weighted ΔP model, these researchers were able to explain the influence of the base rate of the effect on candidates with the same negative contingencies. Note that varying the value of this parameter is necessary if the R–W model is to explain the ordinal pattern of their results. Without this variation, the R–W model is equivalent to the traditional contingency model (Appendix A and Chapman & Robbins, 1990), which is contradicted by such an influence. In contrast, the power PC theory explains Wasserman et al.'s (1983, 1993) pattern of results without any parameters.

Now consider evaluating the generative power of a candidate cause of effects that occur with rates. Whereas a preventive cause decreases the rate of the effect by some proportion of the distance to the lower bound of 0, the rates of the effect produced by generative causes are additive. Compare a situation in which the context is producing the flash at the rate of 0 per second with a situation in which the context is producing it at the rate of .25 per second. A candidate that has the power to produce the flash at a rate of .50 per second would yield an observable difference of +.50 flash per second in rate in both situations. In other words, generative causal powers are additive for effects that occur with rates. Consequently, events that occur with rates have no analogue of (a) the boundary condition for the evaluation of generative power specifying $P(e|\bar{i}) < 1$ and (b) the influence of the base rate of the effect on the magnitude of

defined in terms of probabilities) for cases involving effects that occur with certain rates.

estimated causal power given the same positive contrast. Therefore, the power PC theory cannot be evaluated by Wasserman et al.'s (1993) results regarding positive contingencies.

To evaluate the power PC theory's prediction regarding the influence of $P(e|\bar{i})$ on the judgments of candidate causes with the same positive contingencies, studies using discrete trials are required. I am not aware of studies using discrete trials that are as systematic as Wasserman et al.'s (1983, 1993) for the purpose of evaluating this influence. Some conditions in Allan and Jenkins's (1983) experiments, however, do suggest support for the power PC theory's prediction. A large majority of their conditions, all of which used discrete trials, involved nonnegative contingencies. Among these conditions, only the ones in which the two values of a binary candidate cause consist of an "event" (e.g., a joystick is moved) and a "nonevent" (the joystick is not moved), and the two values of the effect likewise consist of an event (e.g., a dot moving down a computer screen) and a nonevent (the dot remaining in place), seem to provide clearly interpretable results for the purpose of assessing the power PC theory.[12] In these relevant conditions, of the nine pairs of candidate causes that had the same positive contingencies but different $P(e|\bar{i})$, the candidate with a higher $P(e|\bar{i})$ was given a more positive causal rating (i.e., greater causal power) in seven pairs. Some of these differences were large (more than 10 points on a 40-point scale). Only two pairs showed the reversed ordering, and the differences within each pair were relatively small (less than 5 points). Note that the general influence of $P(e|\bar{i})$ on the estimated magnitude of causal power observed here was opposite in direction to that observed for negative contingencies in Wasserman et al.'s (1993) experiments, as predicted by the power PC theory.

Because the R–W model treats positive and negative contingencies symmetrically, if the values assumed for β in that model have the same ordinal values as in Wasserman et al.'s (1993) simulations, this model predicts an influence of $P(e|\bar{i})$ for these positive contingencies opposite in direction to what was generally observed by Allan and Jenkins (1983; see Wasserman et al.'s [1993] simulations of this model). That is, even with variation in parameter values, as long as these values are consistent across studies, the R–W model cannot explain the influence of $P(e|\bar{i})$. In contrast, the power PC theory explains the pattern of results and does so without any parameters.

Summary

The universal focal set in studies of causal induction varying a single candidate cause does reveal causal power according to the power PC theory. These studies show that causal judgments are joint functions of $P(e|\bar{i})$ as well as of ΔP_i, as predicted by Equation 8 of the power PC theory for nonnegative contingencies and as predicted by Equation 14 for nonpositive contingencies. Neither the R–W model nor the traditional contingency model can explain the pattern of results.

Apparent Refutations of the Power PC Theory

A number of phenomena have been interpreted as contradicting the probabilistic contrast model (Cheng & Holyoak, 1995): Given the same values of conditional contrasts, either across trials or across conditions, performance has been shown to vary depending on various factors (Shanks, 1993; Shanks et al., 1996). They might also be considered as contradicting the power PC theory.

These phenomena sometimes involved preasymptotic performance, as in Shanks's (1985a, 1987) studies of acquisition preformance and Dickinson and Burke's (1996) study of the effect of consistency in the pairing of blocking cues and to-be-blocked cues on forward and backward blocking. In the latter study, participants in one of the two conditions (the "varied" pairing condition) were presented with only a single instance of each pairing between a blocked cue and a to-be-blocked cue.

Other times, such phenomena involved components of inference such as conflict resolution or decision making, components that are separable from the process of causal induction. Some studies that reported trial-order effects provided participants with information relevant to causal power that was conflicting or ambiguous across trials, so that participants might not have arrived at a single solution across time (e.g., Chapman, 1991, Experiment 2). For example, in a study by Shanks and Lopez (cited in Shanks, 1995, and Shanks et al., 1996, Experiment 4), participants saw trials in which (a) the combination of cues A and B was followed by an outcome and (b) cue B alone was not followed by the outcome ($AB+$, $B-$) in Stage 1. Then, in Stage 2, they saw trials in which (a) the combination of cues A and C was not followed by the outcome and (b) cue C alone was followed by the outcome ($AC-$, $C+$). Participants also saw a second set of three cues that had an identical pattern of contingencies, except that the patterns for the two stages were reversed. Trial order exerted a strong effect in that participants, at the end of Stage 2, rated A, the critical cue, reliably and substantially less causal than its analog in reverse order.

For this design, the simplest interpretation of information relevant to cue A presented in Stage 1 according to a power analysis is that this cue is causal ($P(e|AB) - P(e|\bar{A}B) = 1$, with the effect occurring at a base rate of 0). Information presented in Stage 2, however, contradicts this interpretation ($P(e|AC) - P(e|\bar{A}C) = -1$, with the effect occurring at a base rate of 1). One way of resolving this conflict is that A changes across phases from producing the effect to preventing the effect. For the analogue of A with the information presented in reverse order, the analogous resolution is that it changes across phases from preventing the effect to producing the effect.

Similarly, the relative validity design (Shanks, 1991, Experiment 3; Wagner et al., 1968; Wasserman, 1990), the other phenomenon mentioned earlier in which the R–W model requires varying values of β, does not allow any unambiguous estimation of causal power (see Melz et al., 1993). Other studies reported that participants exposed to different prior experiences subsequently rated identical candidate causes differently (Williams et al., 1994). In these studies, prior experience encouraged different interpretations of causal power of the same ambiguous situation, so there is no single interpretation across conditions.

[12] Other conditions in their experiments are difficult to interpret because the candidate cause or the effect could each consist of a pair of events (e.g., the joystick is moved to the left or to the right). It is not clear whether participants interpreted each pair of events as defined by the experimenter or as two candidate causes and two effects.

Some trial-order effects involved nonasymptotic performance for a similar reason (Chapman, 1991, Experiment 4). By varying values of α and β, the R–W model explains results from some but not all of these studies involving ambiguous or conflicting estimations of power (see Shanks et al., 1996, for a discussion of some of these phenomena with respect to the R–W model).

A study that involved both the probability and the utility of an outcome has also been interpreted as contradicting the traditional contingency model and the statistical contingency view in general. In a study by Chatlosh et al. (1985, Experiment 2), participants were asked to rate different objective contingencies between an action and an outcome. For some participants, the outcome was understood to be related to monetary gain, whereas, for others, the outcome had no monetary consequences. Chatlosh et al. found that when there was a positive contingency, the ratings were higher in the monetary gain condition than in the neutral one. In contrast, when the contingency was negative, the ratings were more negative in the monetary condition than in the neutral one.

The concepts manipulated in this study—the perceived utility of the outcome and the expected probability of the outcome given certain actions—are distinct in theories of decision making. To see the intuitive difference between these concepts, suppose that whenever Mary asks her father for money, he always gives her 1 cent, whereas whenever she asks her mother, she always gives her 1 dollar. Presumably Mary would direct her requests to her mother rather than to her father.[13] Such a response pattern would not imply that the child has failed to learn that there are deterministic contingencies involved in both cases. Returning to Chatlosh et al.'s (1985) findings, to show that their findings do contradict contingency theories, one would have to demonstrate that it is impossible to develop a dependent measure that is sensitive to this intuitive distinction between utility and expected probability.

Rather than contradicting the power PC theory, the preceding phenomena in fact lie outside its scope. The power PC theory is a computational-level theory of causal induction and, as such, concerns asymptotic performance reflecting this process. These phenomena illustrate other aspects of inference (e.g., a decision-making procedure) that are required for a complete explanation of reasoning performance, both asymptotic and otherwise. It is not my aim in this article to discuss these additional aspects of reasoning, but let me briefly illustrate that these phenomena are consistent with the power PC theory if it is augmented with some plausible assumptions. Acquisition curves, for example, might be explained by adding to the power PC theory an assumption about perceived reliability of causal judgments as a function of sample size (Cheng & Holyoak, 1995) or by adopting a Bayesian process for the assessment of the conditional probabilities in the ΔP metric (Fales & Wasserman, 1992). Likewise, results regarding trial-order effects due to conflicting representations of causal power might be explained by a conflict resolution rule according to which, in the absence of other reasons for resolving a conflict, greater weight is given to the more recent representation.

Implications

Studies of causal induction varying single as well as multiple candidate causes converge in supporting the predictions of the power PC theory. To explain the existence of boundary conditions for interpreting contrast, as revealed in the focal sets that reasoners select from the many possible ones available, the difference between the preferred focal sets for evaluating generative and preventive causal powers, and the different conditions under which a contrast is uninterpretable as an estimate of generative or preventive power, Cheng and Holyoak (1995) had to add auxiliary assumptions to Cheng and Novick's (1990) probabilistic contrast model. These assumptions, most of which have parallels in the principles of experimental design, were based only on intuition. Even with these auxiliary assumptions, Cheng and Holyoak's model cannot explain the influence of the base rate of the effect on the evaluation of candidate causes with the same contingency, let alone the difference between this influence for positive and for negative contingencies. In contrast, all of these phenomena are the mathematical consequences of two equations (Equations 5 and 11), equations that are simple explanations of the probabilistic contrast model by a theory of generative and preventive causal powers. These equations do not contain any parameters and do not require any of Cheng and Holyoak's auxiliary assumptions.

Many of these phenomena—in particular, the asymmetry between generative and preventive causes in its multifarious manifestations—confound the R–W model, even at an ordinal level, despite the use of its parameters. Although this model finds an optimal least-mean-squares solution for a given set of data, such solutions do not necessarily generalize beyond the given set. They often do not. A comparison with the power PC theory suggests that some a priori assumption about the nature of causality is necessary.

Linear Combination Models

An apparent bias that is often described in terms of linear combination models is that frequencies of the effect in the presence of a candidate cause tend to be weighted more than those in its absence (e.g., Schustack & Sternberg, 1981). Adding to the list of phenomena reviewed earlier, this often reported "bias" contradicts the R–W model, even when both of its learning rate parameters are allowed to vary across trials (see Wasserman et al., 1993). In contrast, this bias, just as the influence of $P(e\,|\,\bar{i})$ on noncontingent candidates with the same ΔP_i, follows from the normative power PC theory. Before deriving this tendency from the power PC theory, I first review linear heuristic models with respect to test situations involving a single candidate cause and those involving multiple varying candidate causes.

According to linear heuristic models of causal induction (Arkes & Harkness, 1983; Nisbett & Ross, 1980; Schustack & Sternberg, 1981; Shaklee & Tucker, 1980; Ward & Jenkins, 1965), reasoners base their causal judgments on heuristics that are linear combinations of four frequency variables: a, the frequency of the joint presence of a candidate cause and the effect; b, the frequency of the presence of the candidate cause coupled with the absence of the effect; c, the frequency of the absence

[13] This example is from Michael Waldmann (personal communication, 1993).

of the candidate cause coupled with the presence of the effect; and d, the frequency of the joint absence of a candidate cause and the effect. The weight associated with one or more of these frequencies is zero in some models, which means that causal judgments are based on a subset of the frequencies that have nonzero weight. All variants of such models have a positive weight for a. Schustack and Sternberg's (1981) regression modeling showed that the weights for a and d were positive, whereas those for b and c were negative. Other models assign a zero weight to b, c, or d, but no variant reverses the sign of the weights observed by Schustack and Sternberg.

Empirical and Intuitive Refutations of Linear Models

Linear combination heuristics have been refuted as an account of natural causal induction in studies involving a single candidate cause (Chatlosh et al., 1985; Cheng & Novick, 1991, 1992; Wasserman et al., 1983). Chatlosh et al. (1985) and Wasserman et al. (1983) found that when individual participants' causal ratings were correlated with various judgment rules or heuristics, variants of linear models showed a lower correlation than the ΔP rule. To test linear models and the probabilistic contrast model where their predictions diverge, Cheng and Novick (1992) noted that linear models predict that the prevalence of the effect should influence causal judgments on candidate causes that remain constantly present in the focal set. According to heuristics that have a positive weight for a and a negative weight for b, constantly present candidates should be considered causes when the effect is prevalent (a is large), whereas they should be considered inhibitors when the effect is rare (b is large). For all other existing variants of this class of models (i.e., models that assume a zero weight for b), constant factors should more likely be considered causes when the effect is prevalent than when it is rare. Linear models therefore predict that inhabitants of Edinburgh (where rain is prevalent) should believe that a factor that is constantly present in that environment—such as gravity, car exhaust, or houses—causes rain (the effect), whereas inhabitants of Los Angeles (where rain is rare) should believe that such a factor inhibits rain or should be less likely to believe that such a factor causes rain. Contrary to these predictions, people's intuition and the probabilistic contrast model tell them that gravity is an enabling condition for rain, and car exhaust and houses are irrelevant to rain. Experimental findings confirm that varying the prevalence of the effect had absolutely no impact on judgments of the causal status of candidates that are constantly present (Cheng & Novick, 1991).

Linear models are further refuted by studies involving multiple varying candidate causes, for example, by Fratianne and Cheng (1995) and Park and Cheng's (1995) studies I described earlier. First, note that in Fratianne and Cheng's experiment, not only were the unconditional contrasts equated across the critical candidates B (in Pattern 1) and Y (in Pattern 3), each of the four frequencies used by linear models were also equated across these candidates. This holds under the plausible assumption that participants did not assume different sample sizes when given identical information about candidate causes. Accordingly, linear models, regardless of the weights assigned to the four frequency variables, predict that these candidates should receive the same causal rating. Contradicting this prediction, recall that Y was judged as causal far more confidently than B was.

Second, for Park and Cheng's (1995) experiments, linear models that include c as a parameter (i.e., have a negative weight for c) make the same prediction as the traditional contingency model: The critical candidate B should be rated less causal in the experimental group than in the control group at the end of Phase 1 of both the ceiling and the non-ceiling experiments. Because both candidates A and B were paired with the effect in the experimental group, whereas only B was in the control group, B had a positive c in the experimental group but a zero c in the control group: In the absence of B, the effect sometimes occurred in the experimental group (due to the pairing of A with the effect) but never occurred in the control group. Recall that this prediction was flatly contradicted by Park and Cheng's results.

Third, all linear models predict that, for the control group of Park and Cheng's (1995) ceiling experiment, the causal strength of A should increase from Phase 1 to Phase 2. Recall that in Phase 1 of this group, A was never paired with the effect, whereas in Phase 2, A (in combination with B) was always paired with the effect. The value of a for this candidate should therefore increase from 0 to a positive number. Because all variants of linear models have a positive weight for a, they invariably predict an increase in the causal strength of A across phases. Contrary to this prediction, the mean causal estimate for A showed no sign of such an increase; it in fact showed a small and unreliable decline.

Why Events in Which the Candidate Cause Is Present Are Weighted More

Thus, linear models are inaccurate as models of natural causal induction, both for situations involving a single varying candidate cause and for those involving multiple varying candidate causes. The bias described by such models, however, poses a problem for normative models. It has been observed in many studies that people tend to weight cells a and b more heavily than cells c and d (Anderson & Sheu, 1995; Kao & Wasserman, 1993; Schustack & Sternberg, 1981; Wasserman et al., 1993). For example, Wasserman et al. (1993) found that varying the frequency of the effect in the presence of the candidate cause (a and b) produced a larger range of observed causal ratings than varying this frequency in the absence of the candidate cause (c and d) by an identical amount. This differential weighting is inexplicable by normative contingency models (e.g., Allan, 1980; Cheng & Novick, 1990, 1992; Wasserman et al., 1983). As a result of these and other deviations from the traditional contingency model, Wasserman et al. abandoned it in favor of the R–W model, and Anderson and Sheu (1995) concluded that reasoners use a weighted contingency model. (As mentioned, this differential weighting is also inexplicable by the R–W model.)

I show subsequently that this prediction of linear models that have higher weights for cells a and b than cells c and d coincides with the prediction of the power PC theory. Unlike linear models and the weighted contingency model, however, the power PC theory explains the differential weighting without the use of any parameters (for a normative explanation that adopts a Bayesian approach, see Anderson, 1990).

Why a and b Are Weighted More Than c and d for Generative Causes

First, consider the evaluation of generative causal power. Recall that, according to the power PC theory, the generative nature of a candidate i is assessed by Equation 8. To restrict the range of contrasts to the scope of this equation, assume that $\Delta P_i \geq 0$. Let x represent the numerator of the RHS of this equation (i.e., ΔP_i) and y represent its denominator (i.e., $1 - P(e|\bar{i})$), so that

$$p_i = \frac{P(e|i) - P(e|\bar{i})}{1 - P(e|\bar{i})} = \frac{x}{y}. \qquad (16)$$

By the definitions of x and y, it follows that

$$x = y + P(e|i) - 1. \qquad (17)$$

Note that $P(e|i)$ is estimated by the relative frequency $a/(a + b)$, and $P(e|\bar{i})$ is estimated by the relative frequency $c/(c + d)$.

Increasing contrast. First, consider increasing a nonnegative ΔP_i by some amount z, with $z > 0$. Let us first compare the consequences of doing so by increasing a and by reducing c. Note that to be able to increase a presupposes that $P(e|i) < 1$. Now, because $\Delta P_i \geq 0$, $P(e|i) < 1$ implies that $P(e|\bar{i})$ must also be less than 1. From Equation 17, it can be seen that $P(e|i) < 1$ also implies that $x < y$. In other words, $x/y < 1$. But both x and y are non-negative: $x = \Delta P_i \geq 0$ by assumption, and $y > 0$ because $P(e|\bar{i}) < 1$ and $y = 1 - P(e|\bar{i})$. Therefore, $x/y \geq 0$. In sum, $0 \leq x/y < 1$.

Let me denote the new estimated power of candidate i when a is changed by $p_{i(a)}$. When a is increased,

$$p_{i(a)} = \frac{P(e|i) + z - P(e|\bar{i})}{1 - P(e|\bar{i})} = \frac{x + z}{y}. \qquad (18)$$

When c is decreased by the *same* absolute amount as a is increased, the new estimate due to changing c, which I denote by $p_{i(c)}$, is

$$p_{i(c)} = \frac{P(e|i) - [P(e|\bar{i}) - z]}{1 - [P(e|\bar{i}) - z]} = \frac{x + z}{y + z}. \qquad (19)$$

Because $x/y \geq 0$ and $z > 0$, it follows that

$$\frac{x + z}{y} > \frac{x + z}{y + z}. \qquad (20)$$

That is, $p_{i(a)} > p_{i(c)}$. Now it might be hypothesized that increasing a increases p_i (i.e., $(x + z)/y > x/y$), but reducing c reduces p_i (i.e., $(x + z)/(y + z) < x/y$). On this hypothesis, the change in p_i due to reducing c might have a larger absolute magnitude than that due to increasing a. In fact, increasing ΔP_i by increasing a or reducing c always increases p_i. It is easy to see that, because $z > 0$,

$$\frac{x + z}{y} > \frac{x}{y}. \qquad (21)$$

Now, because $x/y < 1$ (so that z is a larger proportion of x than of y),

$$\frac{x + z}{y + z} > \frac{x}{y}. \qquad (22)$$

Summarizing Equations 20, 21, and 22, we obtain

$$\frac{x + z}{y} > \frac{x + z}{y + z} > \frac{x}{y}. \qquad (23)$$

Thus, the increment in a produces a larger change in p_i than does the decrement in c. That is, a receives more weight than c.

The other two ways of increasing ΔP_i are to reduce b and to increase d. In the preceding argument, because increasing a and reducing b are represented identically as $P(e|i) + z$, and reducing c and increasing d are represented identically as $P(e|\bar{i}) - z$, this argument also applies to a comparison between reducing b and increasing d. (Simply replace "increasing a" in the preceding argument by "reducing b" and replace "reducing c" by "increasing d.") Thus, when a nonnegative ΔP_i is increased an identical amount by decrementing b or incrementing d by the same absolute number, p_i is increased in both cases, and the change in b increases p_i more than does the change in d (i.e., b receives more weight than d).

Reducing contrast. Analogously, consider reducing a nonnegative ΔP_i by amount z. Let us first compare how much reducing a and increasing c changes p_i. As before, $x \geq 0$ by assumption. Now note that to be able to increase c presupposes that $P(e|\bar{i}) < 1$. This implies that $y > 0$. Therefore, $x/y \geq 0$. From Equation 17, it can be seen that $x \leq y$, because $P(e|i) \leq 1$, $P(e|i)$ being a probability. In other words, $x/y \leq 1$. In sum, $0 \leq x/y \leq 1$.

When a is reduced,

$$p_{i(a)} = \frac{x - z}{y}. \qquad (24)$$

When c is increased by the same amount as a is reduced,

$$p_{i(c)} = \frac{x - z}{y - z}. \qquad (25)$$

Because $x/y \geq 0$ and $z > 0$, if the reduction in ΔP_i leaves it within the positive range[14] (i.e., $x > z$), then

$$\frac{x - z}{y} < \frac{x - z}{y - z}. \qquad (26)$$

In other words, $p_{i(a)} < p_{i(c)}$. Analogous to the case of increasing ΔP_i, however, it might be argued that reducing a reduces p_i

[14] If this reduction exactly cancels out ΔP_i (i.e., $x = z$), then $p_{i(a)} = p_{i(c)}$. That is, a receives the same weight as c. And, if the reduction in ΔP_i changes its sign (i.e., $x < z$), then the analysis of the evaluation of preventive causes in the next section applies.

(i.e., $(x - z)/y < x/y$), but increasing c increases p_i (i.e., $(x - z)/(y - z) > x/y$), and that the change in p_i due to changing c has a larger absolute magnitude than that due to changing a. To the contrary, reducing ΔP_i by reducing a or increasing c either reduces or does not change p_i; it never increases it. It is easy to see that, because $z > 0$,

$$\frac{x - z}{y} < \frac{x}{y}. \tag{27}$$

Now, because $x/y \leq 1$ (so that z is at least as large a proportion of x as it is of y);

$$\frac{x - z}{y - z} \leq \frac{x}{y}. \tag{28}$$

Summarizing Equations 26, 27, and 28, we obtain

$$\frac{x - z}{y} < \frac{x - z}{y - z} \leq \frac{x}{y}. \tag{29}$$

Thus, for changes that leave a nonnegative ΔP_i within the positive range, when this ΔP_i is reduced an identical amount by decrementing a or incrementing c by the same absolute number, the change in a reduces p_i more than does the change in c. In summary, for such changes, when a nonnegative ΔP_i is reduced as well as when it is increased, a receives more weight than c.

As explained earlier, the exact same argument applies to a comparison between the effects of increasing b and reducing d by the same absolute amount. Thus, for changes that leave a nonnegative ΔP_i within the positive range, the change in b reduces p_i more than does the change in d. In summary, for such changes, when a nonnegative ΔP_i is reduced as well as when it is increased, b receives more weight than d. (I return later to the case when ΔP_i is reduced to 0 or below.)

Why a and b Are Weighted More Than c and d for Preventive Causes

An analogous analysis of the evaluation of preventive power yields the same differential weighting. Recall that, according to the power PC theory, the preventive nature of a candidate i is assessed by Equation 14. To restrict the range of contrasts to the scope of this equation, assume that $\Delta P_i \leq 0$, which implies that $P(e|i) \leq P(e|\bar{i})$. Now let x represent the numerator of the RHS of this equation (i.e., $-\Delta P_i$) and y represent its denominator (i.e., $P(e|\bar{i})$), so that

$$p_i = \frac{P(e|\bar{i}) - P(e|i)}{P(e|\bar{i})} = \frac{x}{y}. \tag{30}$$

By the definitions of x and y here, it follows that

$$x = y - P(e|i). \tag{31}$$

Increasing a nonpositive contrast (i.e., making it less negative). Consider increasing a nonpositive ΔP_i by some amount z, with $z > 0$ as before. Let us first compare the consequences of doing so by increasing a and by reducing c. Note that to be able to reduce c presupposes that $P(e|\bar{i}) > 0$. That is, $y > 0$. We also know that because $P(e|i) \leq P(e|\bar{i})$, $x \geq 0$. Therefore, $x/y \geq 0$. From Equation 31, it can be seen that because $P(e|i) \geq 0$, $x \leq y$. In other words, $x/y \leq 1$. In sum, $0 \leq x/y \leq 1$.

When a is increased,

$$p_{i(a)} = \frac{P(e|\bar{i}) - [P(e|i) + z]}{P(e|\bar{i})} = \frac{x - z}{y}. \tag{32}$$

When c is decreased by the *same* amount as a is increased,

$$p_{i(c)} = \frac{[P(e|\bar{i}) - z] - P(e|i)}{[P(e|\bar{i}) - z]} = \frac{x - z}{y - z}. \tag{33}$$

As can be seen, the new estimated powers shown in Equations 32 and 33 correspond respectively to those in Equations 24 and 25, the estimates obtained with Equation 8 when a nonnegative ΔP_i is reduced. Also, as for the earlier situation, $0 \leq x/y \leq 1$ here. The previous argument therefore applies here, yielding Equation 29. Thus, for increases in a nonpositive ΔP_i that leave it within the negative range, when this ΔP_i is increased (i.e., when its absolute magnitude is reduced) an identical amount by incrementing a or d or by decrementing c or b, the change in a reduces p_i more than does a change of the same magnitude in c, and the change in b reduces p_i more than does a change of the same magnitude in d. That is, for such changes, a and b receive more weight than c and d.

Reducing a nonpositive contrast (i.e., making it more negative). Now consider reducing a nonpositive ΔP_i by amount z. As before, let us first compare the consequences of doing so by reducing a or increasing c. We know that $x \geq 0$, by assumption, because $\Delta P_i \leq 0$. Now, to be able to reduce a presupposes that $P(e|i) > 0$. Because $\Delta P_i \leq 0$, $P(e|i) > 0$ implies that $y = P(e|\bar{i}) > 0$. Therefore, $x/y \geq 0$. From Equation 31, it follows that $P(e|i) > 0$ also implies $x < y$. In other words, $x/y < 1$. In sum, $0 \leq x/y < 1$.

When a is reduced,

$$p_{i(a)} = \frac{P(e|\bar{i}) - [P(e|i) - z]}{P(e|\bar{i})} = \frac{x + z}{y}. \tag{34}$$

When c is increased by the same amount as a is reduced,

$$p_{i(c)} = \frac{[P(e|\bar{i}) + z] - P(e|i)}{P(e|\bar{i}) + z} = \frac{x + z}{y + z}. \tag{35}$$

As can be seen, the new estimated powers shown in Equations 34 and 35 turn out to correspond respectively to those in Equations 18 and 19, the estimates obtained with Equation 8 when a nonnegative ΔP_i is increased. Also, as for the earlier situation, $0 \leq x/y < 1$. The earlier argument therefore applies here, yielding Equation 23. Thus, when a nonpositive ΔP_i is reduced (i.e., when its absolute magnitude is increased) an identical amount by decrementing a or d or by incrementing c or b, p_i is increased in all cases. Moreover, the change in a increases p_i more than does a change of the same magnitude in c, and the change in b increases p_i more than does a change of the same magnitude in d. That is, a and b receive more weight than c and d.

Let me return now to the case when ΔP_i is exactly cancelled out or when its sign is altered. When ΔP_i is either increased or reduced to 0, a receives the same weight as c, as is evident in Equations 24, 25, 32, and 33. By the same argument, b receives the same weight as d. When the change in ΔP_i alters its sign so that the alternative power equation applies (Equation 8 or 14), the consequences of the total change in ΔP_i on p_i can be assessed by combining the consequences on both sides of 0. Because the derivations for both generative and preventive powers show that a and b receive more weight than c and d (except when ΔP_i is changed to 0), the same differential weighting must hold for a change that crosses from one type of causal power to the other.

Summary

Because linear combination models make predictions that are indisputably refuted by observations of situations that involve either a single candidate cause or multiple candidate causes, these models are implausible as descriptions of natural causal induction. These models, however, capture a robust finding that contradicts "normative" contingency models: Reasoners tend to weigh frequencies of the effect in the presence of a candidate cause (a and c) more than those in its absence (b and d). This differential weighting does not contradict the normative power PC theory; rather, it follows from it. No alternative account can explain both this differential weighting and the various influences of the base rate of the effect on the evaluation of candidates with the same ΔP_i.

The Power PC Theory as a Solution to Problems of the Covariational and Power Views

Summary of the Power PC Theory

Because acquired causal relations are neither directly observable nor deducible, they must be induced. The goal of causal induction is to uncover the causal structure of the world given the input to one's processing system. This task, however, is underconstrained; what appears to be the same pattern of input can stem from either a causal or a noncausal environmental correlate. As I noted, the two previous approaches to causal induction—the covariation and power approaches—each have some fundamental problems. To overcome these problems, I proposed introducing a presumably innate constraint: The environment contains such things as causes that either produce or prevent an effect.[15] Given this constraint, the solution to the problem is to treat the relation between covariation and power as analogous to that between scientists' model or law and their theory of it. Whereas models and laws concern observable entities, theories posit unobservable entities.

This solution, as formalized in the power PC theory, reveals itself in a diverse set of phenomena involving ordinal differences in causal judgments regarding single and multiple generative and preventive candidate causes. These phenomena include the basic influence of contingency (e.g., Baker et al., 1989; Wasserman et al., 1993), the subtle but systematic (and seemingly irrational) influence of the base rate of the effect on the magnitude of causal judgments for candidates with a given negative contingency (Wasserman et al., 1983, 1993), the opposite influence of this base rate on the magnitude of such judgments for candidates with a given positive contingency (Allan & Jenkins, 1983), apparent biases reported in the social psychology literature (Cheng & Novick, 1990), the distinction between causes and enabling conditions (Cheng & Novick, 1991), the distinction between genuine and spurious causes, the distinction between a novel candidate and an irrelevant one (Cheng & Holyoak, 1995), the boundary condition for interpreting generative power as manifested in blocking (Fratianne & Cheng, 1995; Waldmann & Holyoak, 1992) and overexpectation (Park & Cheng, 1995), the boundary condition for interpreting inhibitory power as manifested in the extinction of conditioned inhibition (e.g., Williams, 1995; Yarlas et al., 1995), the asymmetry between these boundary conditions, retrospective changes in causal judgments (e.g., Chapman, 1991; Yarlas et al., 1995), and the apparently irrational tendency to weigh information regarding the presence of a candidate cause more than that regarding its absence (e.g., Anderson & Sheu, 1995; Schustack & Sternberg, 1981; Wasserman et al., 1993). Whereas the power PC theory provides a unified and parameter-free explanation for these phenomena, every alternative model of causal induction fails to explain many of them.

At the same time that the power PC theory explains many apparent biases, this theory provides an explanation of the principles of experimental design, such as the use of control groups, random assignment, avoidance of ceiling effects, and the requirement that all extraneous variables be held constant between the experimental and control groups. Unlike Cheng and Holyoak's (1995) model, which enlists some of these principles as assumptions, the power PC theory provides a unified explanation and justification for these principles.

I assume that the power PC theory describes an innate component of the process of causal induction. In this theory, the explanation of a model by a theory and the constraint mentioned earlier are a priori assumptions; without them, causal induction cannot begin. These assumptions seem to be necessary for explaining the critical findings involving causal reasoning in humans, because these findings are inexplicable by alternative models. However, these findings (e.g., the asymmetry in boundary conditions for interpreting generative and preventive power) often have parallels in classical conditioning studies using laboratory animals, findings that are equally inexplicable by alternative models. Now, whereas it might be argued that college students (participants in the human studies) have attended many science classes in which to learn the idea of explaining a model by a theory, meek laboratory animals are seldom offered the opportunity. The a priori assumptions in the power PC theory must therefore be innate.

My approach has its roots not only in the works of Hume (1739/1987) and Kant (1781/1965) but also in more recent work. A number of researchers have proposed that induction is normative (Kelley, 1967, 1973; Peterson & Beach, 1967). Others have specifically interpreted contrast in terms of causal power (Baker et al., 1989; Cheng & Novick, 1992; Dickinson

[15] Another such constraint that I discussed—that causes and effects can occur in the form of various types of variables (e.g., a discrete effect can occur in discrete entities or in continuous time)—is not specific to the process of causal induction. There are no doubt other innate constraints, notably domain-specific ones. This article does not discuss the latter constraints because I would like to see what can be explained by minimal assumptions about innateness.

et al., 1984; Waldmann & Holyoak, 1992) or more generally interpreted covariation as the measurement of, or evidence for, causal power (Ahn et al., 1995; Cartwright, 1989; Waldmann, 1996). The contributions of the power PC theory are that (a) it is the first example of a formal theory of a psychological model, (b) it explains a diverse range of phenomena regarding causal induction, and (c) it solves some basic problems afflicting the covariation and power views of causal induction.

A Solution to Problems of the Covariational and Power Views

The relation between a model and a theory of the model makes it possible to pinpoint the conditions under which covariation implies causation, thus overcoming the most fundamental problem afflicting the covariation view. This relation provides a justification for the leap from covariation to causation: If one is willing to assume that there are such things as causes that either produce or prevent an effect, then, under specific conditions, covariation reveals causal power. The general inequality between covariation and causation is a problem that pervades all covariational models of causality, associationist or statistical. Without a causal power theory, it is difficult to see how any covariation model can free itself of the chains that bind the interpretation of even formal statistical covariation.

At the same time, the power PC theory motivates a separate assumption typically made by covariation models: that causes are temporally prior to their effects (see Pearl, 1996; Waldmann & Holyoak, 1992). Covariation between two types of events has no inherent temporal order. It is possible, for example, to form an association from the effect to the cause. But if reasoners have the intuitive notion that causes produce (or prevent) the effect, it follows that the cause must precede the effect (even if only by an infinitesimal amount) because a cause must exist before it can produce any consequences. The cause should therefore precede the effect at least in theory, if not by measurable time. The power PC theory provides a coherent link between covariation and temporal priority.

The assumption that people have a causal power theory of their covariation model not only solves the preceding problems afflicting the covariation view, but also solves the two problems afflicting the power view. Recall that this view previously has never presented a solution to the problem of causal induction; it has never specified a mapping between the input and the output of the causal induction process. Unlike previous variants of the power view, the power PC theory specifies how the final output, which is an estimate of the causal power of a specific candidate cause, is computed from the input, which consists of the input to the covariation process. This input is solely restricted to observable events: the presence and absence of the candidate causes and of the effect. The power PC theory thereby honors Hume's indisputable point that causal relations are not explicit in one's sensory input, at the same time that it specifies the a priori causal knowledge that interacts with the sensory-based input.

Also recall that the power view has appeared to be circular, implying that people do not learn that a relation is causal unless they already understand it to be causal. The power PC theory removes this circularity. The a priori causal knowledge assumed is general rather than specific. This theory specifies how a general notion of causes producing or stopping an effect can interact with observable information to yield a theory of covariation, a theory that allows specific causal powers to be inferred. According to this theory, the causal power of a candidate cause can be assessed without prior knowledge about itself, or even the identity of alternative causes. All that is required as input is observable information sufficient to separate the causal power of the candidate from that of alternative causes.

I have not touched on the important issue of how prior domain-specific causal knowledge (whether innate or learned) regarding superordinate kinds influences subsequent causal judgments. Even within the issue of how reasoners come to know that one thing causes another, I have focused only on (a) effects and candidate causes that are clearly defined and that can be represented in terms of probabilities and (b) simple causes that influence the occurrence of an effect independently of background causes within a context. For such situations, however, the theory I proposed presents a theoretical solution to the problem of causal induction first posed by Hume more than two and a half centuries ago. Moreover, the fact that this theory provides a simple explanation for a diverse set of phenomena regarding human reasoning and Pavlovian conditioning suggests that it *is* the solution adopted biologically by humans and perhaps other animals.

References

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

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54,* 299–352.

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

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology, 34,* 1–11.

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

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108,* 441–485.

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.

Arkes, H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General, 112,* 117–135.

Baker, A. G., Berbrier, M. W., & Vallée-Tourangeau, F. (1989). Judgments of a 2 × 2 contingency table: Sequential processing and the learning curve. *Quarterly Journal of Experimental Psychology, 41B,* 65–97.

Baker, A. G., & Mackintosh, N. J. (1976). Learned irrelevance and learned helplessness: Rats learn that stimuli, reinforcers, and responses are uncorrelated. *Journal of Experimental Psychology: Animal Behavior Processes, 2,* 130–141.

Baker, A. G., & Mackintosh, N. J. (1977). Excitatory and inhibitory conditioning following uncorrelated presentations of the CS and UCS. *Animal Learning and Behavior, 5,* 315–319.

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: The presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 414–432.

Baker, A. G., Murphy, R. A., & Vallée-Tourangeau, F. (1996). Associative and normative accounts of causal induction: Reacting to versus understanding a cause. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 1–46). San Diego, CA: Academic Press.

Baron, J. (1994). *Thinking and deciding* (2nd ed.). New York: Cambridge University Press.

Best, M. R., Dunn, D. P., Batson, J. D., Meachum, C. L., & Nash, S. M. (1985). Extinguishing conditioned inhibition in flavor-aversion learning: Effects of repeated testing and extinction of the excitatory element. *Quarterly Journal of Experimental Psychology, 37B,* 359–378.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

Cartwright, N. (1989). *Nature's capacities and their measurement.* Oxford, England: Clarendon Press.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 837–854.

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

Chatlosh, D. L., Neunaber, D. J., & Wasserman, E. A. (1985). Response-outcome contingency: Behavioral and judgmental effects of appetitive and aversive outcomes with college students. *Learning and Motivation, 16,* 1–34.

Cheng, P. W. (1993). Separating causal laws from casual facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 30, pp. 215–264). New York: Academic Press.

Cheng, P. W., & Fratianne, A. (1995, November). *Implicit versus explicit causal reasoning.* Paper presented at poster session at the 36th annual meeting of the Psychonomic Society, Los Angeles, CA.

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

Cheng, P. W., & Lien, Y. (1995). The role of coherence in distinguishing between genuine and spurious causes. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 463–494). Oxford, England: Oxford University 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. (1991). Causes versus enabling conditions. *Cognition, 40,* 83–120.

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

Cheng, P. W., Park, J., Yarlas, A. S., & Holyoak, K. J. (1996). A causal-power theory of focal sets. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 313–357). San Diego, CA: Academic Press.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology, 49B,* 60–80.

Dickinson, A., & Shanks, D. R. (1986). The role of selective attribution in causality judgment. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 94–126). Brighton, England: Harvester Press.

Dickinson, A., Shanks, D. R., & Evenden, J. L. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A,* 29–50.

Downing, C. J., Sternberg, R. J., & Ross, B. H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General, 114,* 239–263.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3–19.

Fales, E., & Wasserman, E. A. (1992). Causal knowledge: What can psychology teach philosophers? *Journal of Mind and Behavior, 13,* 1–27.

Feller, W. (1957). *An introduction to probability theory and its applications* (2nd ed.). New York: Wiley. (Original work published 1950)

Fratianne, A., & Cheng, P. W. (1995). *Assessing causal relations by dynamic hypothesis testing.* Manuscript submitted for publication.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Gluck, M., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Goodman, N. (1983). The new riddle of induction. In *Fact, fiction, and forecast* (4th ed., pp. 59–83). Cambridge, MA: Harvard University Press. (Original work published 1954)

Hallam, S. C., Matzel, L. D., Sloat, J. S., & Miller, R. R. (1990). Excitation and inhibition as a function of posttraining extinction of the excitatory cue used in Pavlovian inhibition training. *Learning and Motivation, 21,* 59–84.

Harré, R., & Madden, E. H. (1975). *Causal powers: A theory of natural necessity.* Totowa, NJ: Rowman & Littlefield.

Hart, H. L., & Honoré, A. M. (1985). *Causation in the law* (2nd ed.). Oxford, England: Oxford University Press. (Original work published 1959)

Hume, D. (1987). *A treatise of human nature* (2nd ed.). Oxford, England: Clarendon Press. (Original work published 1739)

Jaspars, J. M. F., Hewstone, M. R. C., & Fincham, F. D. (1983). Attribution theory and research: The state of the art. In J. M. F. Jaspars, F. D. Fincham, & M. R. C. Hewstone (Eds.), *Attribution theory: Essays and experiments* (pp. 3–36). San Diego, CA: Academic Press.

Jenkins, H., & Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 7,* 1–17.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior, 1967: Aversive stimulation* (pp. 9–31). Coral Gables, FL: University of Miami Press.

Kant, I. (1965). *Critique of pure reason.* London: Macmillan. (Original work published 1781)

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1363–1386.

Kaplan, P. S., & Hearst, E. (1985). Excitation, inhibition, and context: Studies of extinction and reinstatement. In P. D. Balsam & A. Tomie (Eds.), *Context and learning* (pp. 195–224). Hillsdale, NJ: Erlbaum.

Kasprow, W. J., Schachtman, T. R., & Miller, R. L. (1987). The comparator hypothesis of conditioned response generation: Manifest conditioned excitation and inhibition as a function of relative excitatory strengths of CS and conditioning context at the time of testing. *Journal of Experimental Psychology: Animal Behavior Processes, 13,* 395–406.

Kelley, H. H. (1967). Attribution theory in social psychology. In D.

Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 192–238). Lincoln: University of Nebraska Press.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28,* 107–128.

Kremer, E. F. (1971). Truly random and traditional control procedures in CER conditioning in the rat. *Journal of Comparative and Physiological Psychology, 76,* 441–448.

Kremer, E. F. (1978). The Rescorla-Wagner model: Losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 4,* 22–36.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science, 5,* 3–36.

Lysle, D. T., & Fowler, H. (1985). Inhibition as a "slave" process: Deactivation of conditioned inhibition through extinction of conditioned excitation. *Journal of Experimental Psychology: Animal Behavior Processes, 11,* 71–94.

Mackie, J. L. (1974). *The cement of the universe: A study of causation.* Oxford, England: Clarendon Press.

Mackintosh, N. J. (1983). *Conditioning and associative learning.* Oxford, England: Clarendon Press.

Marr, D. (1982). *Vision.* New York: Freeman.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

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

Michotte, A. E. (1963). *The perception of causality.* New York: Basic Books. (Original work published 1946)

Mill, J. S. (1973). A system of logic ratiocinative and inducive. In J. M. Robson (Ed.), *Collected works of John Stuart Mill* (Vols. 7, 8). Toronto, Ontario, Canada: University of Toronto Press. (Original work published 1843)

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla–Wagner model. *Psychological Bulletin, 117,* 363–386.

Miller, R. R., & Schachtman, T. R. (1985). Conditioning context as an associative baseline: Implications for response generation and the nature of conditioned inhibition. In R. R. Miller & N. E. Spear (Eds.), *Information processing in animals: Conditioned inhibition* (pp. 51–88). Hillsdale, NJ: Erlbaum.

Morris, W. M., & Larrick, R. (1995). When one cause casts doubts on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102,* 331–355.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and short-comings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Novick, L. R., Fratianne, A., & Cheng, P. W. (1992). Knowledge-based assumptions in causal attribution. *Social Cognition, 10,* 299–333.

Park, J.-Y., & Cheng, P. W. (1995). *Boundary conditions on "overexpectation" in causal learning with discrete trials: A test of the power PC theory.* Manuscript in preparation.

Pavlov, I. P. (1927). *Conditioned reflexes.* New York: Dover.

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

Pearl, J. (1996). Structural and probabilistic causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 393–435). San Diego, CA: Academic Press.

Peterson, C. R., & Beach, L. R. (1967). Man as intuitive statistician. *Psychological Bulletin, 68,* 29–46.

Price, P. C., & Yates, J. F. (1993). Judgmental overshadowing: Further evidence of cue interaction in contingency judgment. *Memory and Cognition, 21,* 561–572.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66,* 1–5.

Rescorla, R. A. (1969). Conditioned inhibition of fear resulting from CS-US contingencies. *Journal of Comparative and Physiological Psychology, 67,* 504–509.

Rescorla, R. A. (1970). Reduction in the effectiveness of reinforcement after prior excitatory conditioning. *Learning and Motivation, 1,* 372–381.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43,* 151–160.

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

Rumelhart, D. E., & McClelland, J. L. (1986). On learning past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Salmon, W. C. (1965). The status of prior probabilities in statistical explanation. *Philosophy of Science, 32,* 137–146.

Salmon, W. C. (1980). Probabilistic causality. *Pacific Philosophical Quarterly, 61,* 50–74.

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

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory and Cognition, 8,* 459–467.

Shanks, D. R. (1985a). Continuous monitoring of human contingency judgment across trials. *Memory and Cognition, 13,* 158–167.

Shanks, D. R. (1985b). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology, 37B,* 1–21.

Shanks, D. R. (1987). Acquisition functions in causality judgment. *Learning and Motivation, 18,* 147–166.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

Shanks, D. R. (1993). Associative versus contingency accounts of category learning: Reply to Melz, Cheng, Holyoak, and Waldmann (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1411–1423.

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

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

Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 265–312). San Diego, CA: Academic Press.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47* (Serial No. 1).

Siegel, S., & Allan, L. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin and Review, 3,* 314–321.

Spellman, B. A. (1996a). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7,* 337–342.

Spellman, B. A. (1996b). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 167–207). San Diego, CA: Academic Press.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam: North-Holland.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88,* 135–170.

Taylor, R. (1967). Causation. In P. Edwards (Ed.), *The encyclopedia of philosophy* (Vol. 2, pp. 56–66). New York: Macmillan.

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology, 76,* 171–180.

Waldmann, M. R., (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 47–88). San Diego, CA: Academic Press.

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

Ward, W., & Jenkins, H. (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 structure of the environment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 27–82). San Diego, CA: Academic Press.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation, 14,* 406–432.

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

Wasserman, E. A., Kao, S.-F., Van Hamme, L., Katagiri, M., & Young, M. (1996). Causation and association. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 208–264). San Diego, CA: Academic Press.

White, P. A. (1989). A theory of causal processing. *British Journal of Psychology, 80,* 431–454.

White, P. A. (1995). Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory and Cognition, 23,* 243–254.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Part 4,* 96–104.

Williams, D. A. (1995). Forms of inhibition in animal and human learning. *Journal of Experimental Psychology: Animal Behavior Processes, 21,* 129–142.

Williams, D. A. (1996). A comparative analysis of negative contingency learning in humans and nonhumans. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 89–132). San Diego, CA: Academic Press.

Williams, D. A., & Docking, G. L. (1995). Associative and normative accounts of negative transfer. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48A,* 976–998.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 694–709.

Yarlas, A. S., Cheng, P. W., & Holyoak, K. J. (1995). Alternative approaches to causal induction: The probabilistic contrast versus the Rescorla–Wagner model. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 431–436). Hillsdale, NJ: Erlbaum.

Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology, 86,* 837–845.

(Appendixes follow on next page)

Appendix A

Deriving Asymptotic Weights for Designs for Which the R–W Model Does and Does Not Compute Conditional Contrasts

Deriving Asymptotic Weights

The goal of this appendix is to characterize the R–W model (1972) at Marr's (1982) computational level, that is, to find the mathematical function characterizing the model's asymptotic behavior. Because my derivation concerns asymptotic weights, I first describe a method for deriving such weights (Melz et al., 1993). To obtain the asymptotic weights of a two-layered network that updates the weights of its links according to the R–W model, first note the equivalence between the R–W learning rule and the least mean squares rule of Widrow and Hoff (1960; see Sutton & Barto, 1981). The equivalence implies that the R–W rule implements an iterative algorithm for computing the solution to a set of linear equations defined by the set of stimulus–response patterns presented to the network (Widrow & Hoff, 1960). A *pattern* is a configuration of stimuli and a response deterministically describing a set of trials. If the input patterns are linearly independent, then the R–W rule will discover a unique solution. Even if the input patterns are not linearly independent, the network will still converge provided that the learning rate is sufficiently small and that the various patterns occur with sufficient frequency in the input sequence. The network will converge so as to minimize the sum of the squared errors over the patterns. That is, the equation

$$\mathbf{E} = \sum_p \pi_p \iota_p \left(\lambda_p - \sum_i V_{pij} \right)^2 \tag{A1}$$

will be minimized, where p is the index for a particular stimulus–response pattern, π_p is the frequency of pattern p, ι_p is the learning rate associated with pattern p (τ_j and γ_j, respectively, for the presence and the absence of outcome j), λ_p is the actual outcome for the pattern (usually represented as 0 when j is absent and as 1 when j is present), and $\sum_i V_{pij}$ is the predicted outcome for the pattern, which is equal to the sum of the weights V_i associated with every cue i occurring in the pattern. τ_j is the product of the rate parameters α_i and β_j in Equation 15 for trials on which j is present; γ_j is this product for trials on which j is absent. If $\tau_j = \gamma_j$, the ι_p term may be omitted from the equation. I assume that $\tau_j = \gamma_j$ in the rest of this appendix.

Thus, the asymptotic weights of a network, according to the R–W model, can be calculated analytically by minimizing the sum of the squared errors given by Equation A1. This minimum value may be obtained by setting the partial derivatives with respect to each weight to 0 and solving the resulting set of equations.

Asymptotic Weights for Designs for Which the R–W Model Computes Conditional Contrasts

Design With Two Cues

To illustrate the instantiation of Equation A1, first consider a simple contingency design involving two candidate cues, Cue 1 and Cue 2. Cue 1 is the constant context assumed by applications of the R–W model (see Rescorla & Wagner, 1972), and Cue 2 is a varying cue. Because one of my goals in this section is to provide a preview of the argument used in the more general derivation of the relation between R–W and conditional contrasts, I do not apply Equation A1 here in the most straightforward manner.

In this design, there are two patterns of trials on which Cue 1 is the only cue present. Assume that there are a total of k_1 such trials, with the outcome j occurring ($\lambda_1 = 1$) on π_1 trials (one pattern) and the outcome not occurring ($\lambda_1 = 0$) on $k_1 - \pi_1$ trials (the other pattern). According to Equation A1, the error from these two patterns, denoted as $\mathbf{E}_1$, is as follows:

$$\begin{aligned}\mathbf{E}_1 &= \pi_1(1 - V_1)^2 + (k_1 - \pi_1)(0 - V_1)^2 \\ &= k_1 V_1^2 - 2\pi_1 V_1 + \pi_1. \end{aligned} \tag{A2}$$

Therefore, the partial derivative of $\mathbf{E}_1$ with respect to V_1 is

$$\frac{\partial \mathbf{E}_1}{\partial V_1} = 2k_1 V_1 - 2\pi_1. \tag{A3}$$

Setting Equation A3 to 0 yields

$$V_1 = \frac{\pi_1}{k_1}. \tag{A4}$$

That is, considering in isolation the trials on which a single cue (Cue 1 in this case) is present, the associative strength between that cue and the outcome (i.e., the weight of the link) is equal to the relative frequency of trials on which the outcome occurs in the presence of that cue.

Effect of adding a combination containing an additional cue on the partial derivative of the total error **E** *with respect to* V_1. Now consider two additional patterns in which Cue 1 and Cue 2 are both present. Assume that there are a total of k_2 such trials, with the outcome j occurring on π_2 trials and not occurring on $k_2 - \pi_2$ trials. According to Equation A1, for a network with these four patterns,

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_{+2}, \tag{A5}$$

where **E** denotes the total error and $\mathbf{E}_{+2}$ denotes the error due to the two patterns in which Cue 2 is included in addition to Cue 1.

We know that

$$\begin{aligned}\mathbf{E}_{+2} &= \pi_2[1 - (V_1 + V_2)]^2 + (k_2 - \pi_2)[0 - (V_1 + V_2)]^2 \\ &= k_2 V_1^2 + k_2 V_2^2 + 2k_2 V_1 V_2 - 2\pi_2 V_1 - 2\pi_2 V_2 + \pi_2. \end{aligned} \tag{A6}$$

It follows that the partial derivatives of $\mathbf{E}_{+2}$ with respect to V_1 and V_2 are, respectively,

$$\frac{\partial \mathbf{E}_{+2}}{\partial V_1} = 2k_2 V_1 + 2k_2 V_2 - 2\pi_2 \tag{A7}$$

and

$$\frac{\partial \mathbf{E}_{+2}}{\partial V_2} = 2k_2 V_1 + 2k_2 V_2 - 2\pi_2. \tag{A8}$$

Because **E** is the sum of $\mathbf{E}_1$ and $\mathbf{E}_{+2}$ (Equation A5), it follows that

$$\frac{\partial \mathbf{E}}{\partial V_1} = \frac{\partial \mathbf{E}_1}{\partial V_1} + \frac{\partial \mathbf{E}_{+2}}{\partial V_1}. \tag{A9}$$

From Equations A7 and A8, we see that

$$\frac{\partial \mathbf{E}_{+2}}{\partial V_1} = \frac{\partial \mathbf{E}_{+2}}{\partial V_2}. \tag{A10}$$

But $\partial \mathbf{E}_{+2}/\partial V_2$ is equal to the partial derivative of the total error **E** with respect to Cue 2 (i.e., $\partial\mathbf{E}_{+2}/\partial V_2 = \partial\mathbf{E}/\partial V_2$). This is because $\partial\mathbf{E}_1/\partial V_2 = 0$ as a result of $\mathbf{E}_1$ not containing any V_2 terms. Like the partial derivative of $\mathbf{E}$ with respect to all other cues, $\partial\mathbf{E}/\partial V_2$ is set to 0 to minimize the sum of squared errors. Now, because $\partial\mathbf{E}_{+2}/\partial V_1 = \partial\mathbf{E}_{+2}/\partial V_2 = \partial\mathbf{E}/\partial V_2 = 0$,

$$\frac{\partial\mathbf{E}}{\partial V_1} = \frac{\partial\mathbf{E}_1}{\partial V_1} + \frac{\partial\mathbf{E}_{+2}}{\partial V_1} = \frac{\partial\mathbf{E}_1}{\partial V_1}. \tag{A11}$$

That is, the partial derivative of the total error $\mathbf{E}$ with respect to V_1 remains unchanged by the addition of the combination containing V_2 in addition to V_1; therefore, as for the previous case in which Cue 1 occurs alone, $V_1 = \pi_1/k_1$.

Solving for V_2. To solve for V_2, we set Equation A8 to 0, obtaining

$$k_2V_1 + k_2V_2 = \pi_2. \tag{A12}$$

From Equations A4 and A12, it follows that

$$\begin{aligned} V_2 &= \frac{\pi_2}{k_2} - V_1 \\ &= \frac{\pi_2}{k_2} - \frac{\pi_1}{k_1}. \end{aligned} \tag{A13}$$

Recall that π_2/k_2 is the relative frequency of trials on which the outcome occurs in the presence of Cue 1 and Cue 2, and π_1/k_1 is the relative frequency of trials on which the outcome occurs in the presence of Cue 1 and the absence of Cue 2. Therefore, V_2, the strength of the link from Cue 2 to outcome j, estimates $P(j|\text{Cue }2\cdot\text{Cue }1) - P(j|\overline{\text{Cue }2}\cdot\text{Cue }1)$, the contrast for Cue 2 with respect to outcome j conditional on the presence of Cue 1. (A dot between the names of cues in the contrast denotes "and." This notation is omitted when a single letter represents a cue.) A different derivation of the same result was presented by Chapman and Robbins (1990).

Design With n Nested Cues

To generalize the preceding result, I consider a design with n cues (n is an integer greater than 1) in which every combination of cues except the one with a single cue can be characterized as a proper superset of all sets with fewer cues. I refer to such cue combinations as *nested.* (All cues that are always present or absent together are treated as a single composite cue. Cues that are never presented in combination with any of the n cues do not affect the weights of these cues[A1] and are not considered part of the design for my purpose here.) If I denote cues by letters and combinations by sequences of letters, then an example of a nested set of combinations would be a, ab, abc. An example of a nonnested set would be a, ab, bc, where the bc combination is not a superset of the smaller set a. I show subsequently that, for any combination with multiple cues in a nested set, the strength of the cue in it that does not belong to the next smaller combination is equal to the contrast for that cue conditional on the presence of the rest of the cues in the larger combination.

Effect of adding a combination that contains a novel cue. First, consider forming a new combination by adding novel Cue n to a combination with $n - 1$ distinct cues ordered from Cue 1 to Cue $n - 1$. Then add this new combination to the set of all combinations containing any cue from Cue 1 to Cue $n - 1$. Let k_n be the total number of trials with all n cues present, and let π_n be the number of such trials on which the outcome occurs. By adding this new combination, two patterns are therefore added: On π_n trials, all n cues are present, and the outcome occurs; on $k_n - \pi_n$ trials, all n cues are again present, but the outcome does not occur. The sum of the additional squared error terms due to these two patterns, $\mathbf{E}_{+n}$, is

$$\begin{aligned} \mathbf{E}_{-n} &= \pi_n\left(1 - \sum_{i=1}^{n} V_i\right)^2 + (k_n - \pi_n)\left(0 - \sum_{i=1}^{n} V_i\right)^2 \\ &= k_n \sum_{i=1}^{n} (V_i^2) + 2k_n \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} V_iV_j - 2\pi_n \sum_{i=1}^{n} V_i + \pi_n. \end{aligned} \tag{A14}$$

(The double summation term in Equation A14 collects all V_iV_j terms for which $i \neq j$. I separate the V_iV_j terms from the V_i^2 terms for the purpose of obtaining their partial derivatives separately.) Note two implications of Equation A14. First, it implies that the partial derivative of $\mathbf{E}_{+n}$ with respect to V_n is

$$\frac{\partial\mathbf{E}_{+n}}{\partial V_n} = 2k_nV_n + 2k_n \sum_{i=1}^{n-1} V_i - 2\pi_n. \tag{A15}$$

But $\partial\mathbf{E}_{+n}/\partial V_n$ is also the partial derivative of the total error $\mathbf{E}$ with respect to V_n (i.e., $\partial\mathbf{E}_{+n}/\partial V_n = \partial\mathbf{E}/\partial V_n$). This follows because $\mathbf{E}_{n-1}$, the total error for all patterns in the design except those containing cue n, does not contain any V_n terms, implying that $\partial\mathbf{E}_{n-1}/\partial V_n = 0$. (By definition, the added patterns are the only two patterns in the entire design in which cue n appears.) Setting $\partial\mathbf{E}/\partial V_n$ (which is equal to $\partial\mathbf{E}_{+n}/\partial V_n$) to 0, we obtain

$$V_n + \sum_{i=1}^{n-1} V_i = \frac{\pi_n}{k_n}. \tag{A16}$$

That is, when a combination contains novel cue n and $n - 1$ other cues that appear in other combinations, the sum of the strengths of all n cues is equal to the relative frequency of the outcome for that combination of cues. Therefore,

$$V_n = \frac{\pi_n}{k_n} - \sum_{i=1}^{n-1} V_i. \tag{A17}$$

Second, Equation A14 also implies that

$$\frac{\partial\mathbf{E}_{+n}}{\partial V_n} = \frac{\partial\mathbf{E}_{+n}}{\partial V_i} \quad \text{for} \quad i = 1, 2, \ldots, n - 1. \tag{A18}$$

This is so because all n cues are present in the two additional patterns and are not differentiable from each other with respect to these patterns alone. Because $\partial\mathbf{E}_{+n}/\partial V_n = \partial\mathbf{E}/\partial V_n = 0$, it follows that

$$\frac{\partial\mathbf{E}_{+n}}{\partial V_i} = 0 \quad \text{for} \quad i = 1, 2, \ldots, n - 1. \tag{A19}$$

[A1] Applying Equation A1, we see that the sum of the additional squared error terms due to adding a separate cue s, $\mathbf{E}_{+s}$, does not contain any term involving any of the n cues. Its partial derivative with respect to any of the n cues is therefore 0, implying that adding cue s does not affect the partial derivative of the total error $\mathbf{E}$ with respect to any of them. In other words, the set of equations for deriving the asymptotic weights of the n cues will remain unchanged.

(*Appendixes continue on next page*)

Thus, adding the combination containing cue n does not change the partial derivative of the total error $\mathbf{E}$ with respect to any of the other cues (i.e., $\partial\mathbf{E}/\partial V_i = \partial\mathbf{E}_{n-1}/\partial V_i$ for $i = 1, 2, \ldots, n-1$) and, hence, does not affect the asymptotic weights of any of the other cues. In other words, adding a combination that contains a cue that does not appear in any other combination does not change the asymptotic weights of the cues in the other combinations.

Iterating solution. Because the addition of the combination that contains novel cue n does not affect the asymptotic weights of any of the other cues, this combination can be ignored. Now suppose that the remaining set of combinations formed by the $n-1$ cues is nested, so that the smallest combination consists of Cue 1, the next smallest consists of Cue 1 and Cue 2, and so on, with the largest consisting of Cue 1 to Cue $n-1$. The same reasoning that yields Equation A16 then applies to every outermost combination in the remaining nested set as one iteratively peels off the previously outermost combination. In sum, for any cue x in the nested set, where $n \geq x \geq 1$,

$$\sum_{i=1}^{x} V_i = \frac{\pi_x}{k_x}. \tag{A20}$$

Applying Equation A20 to the combination that contains all cues from Cue 1 to Cue $n-1$ yields

$$\sum_{i=1}^{n-1} V_i = \frac{\pi_{n-1}}{k_{n-1}}. \tag{A21}$$

Making use of Equation A21 to solve for V_n in Equation A17, we obtain

$$V_n = \frac{\pi_n}{k_n} - \frac{\pi_{n-1}}{k_{n-1}}. \tag{A22}$$

That is, V_n, the weight of the link from cue n to outcome j, estimates $P(j|\text{Cue } 1 \cdot \text{Cue } 2 \ldots \cdot \text{Cue } n-1 \cdot \text{Cue } n) - P(j|\text{Cue } 1 \cdot \text{Cue } 2 \ldots \cdot \text{Cue } n-1 \cdot \overline{\text{Cue } n})$, the contrast for cue n conditional on the presence of the other $n-1$ cues.

Because peeling off outer combinations containing Cue n or Cue $n-1$ does not affect the asymptotic weights of any of the other cues, the preceding derivation can be applied to any cue x and cue $x-1$ in the nested set for $n \geq x > 1$, yielding the general result

$$V_x = \frac{\pi_x}{k_x} - \frac{\pi_{x-1}}{k_{x-1}}. \tag{A23}$$

That is, V_x, the strength of the link from any cue x in the nested set to outcome j, estimates the contrast for that cue conditional on the presence of the $x-1$ cues in the next smaller combination in the set.

In sum, in a design with multiple cues, if every combination of cues except the one with a single cue can be characterized as a proper superset of all sets with fewer cues, then the strengths of the cues in each combination sum to the relative frequency of the outcome given that combination (Equation A20). These frequencies estimate the corresponding conditional probabilities. Because the strengths of the cues are additive in the R–W model, it follows that, for any combination with multiple cues, the strength of the cue in it that does not belong to the next smaller combination is equal to the contrast for that cue conditional on the presence of the cues in the smaller combination (i.e., the rest of the cues in the larger combination; Equation A23).

Generalizing to Partially Overlapping Combinations

My derivation of the asymptotic weights of the R–W model for a nested design can be readily generalized to designs involving partially overlapping cue combinations. By partially overlapping, I mean combinations that share some cues but that each have distinctive cues (i.e., they are neither disjoint nor a superset or subset of each other). For such designs, if for every pair of combinations that partially overlap with each other, all supersets of one combination (including itself) share the same intersection with the other combination, and this intersection occurs as a separate combination, then the R–W model still computes conditional contrasts asymptotically as specified earlier. Such sets share a common "trunk" of a set of cue combinations that are nested (including the trivial case of a single combination) but then branch out in different nested cue combinations involving additions of cues that are disjoint across branches. An example of a design that satisfies this condition is *a*, *ab*, and *ac*. In this design, the combinations *ab* and *ac* partially overlap. They have distinctive cues *b* and *c* but share a common trunk *a* that occurs as a separate combination. Another example is the design *a*, *ab*, *abc*, *ad*, and *ade*. Consider combinations *ab* and *ad* in this design. They partially overlap with each other, with *a* as their intersection, and this intersection occurs as a separate combination. All supersets of *ab* (i.e., *ab* and *abc*) share this intersection with *ad*. Likewise, all supersets of *ad* (i.e., *ad* and *ade*) share this intersection with *ab*. As I show subsequently, the R–W model still computes conditional contrasts for such sets.

First, consider n cues ordered from Cue 1 to Cue n, the combinations of which form A, the nested set considered earlier with no partially overlapping combinations. Second, consider a trunk that consists of set B, a subset of A for which k is the cue unique to the largest combination, $n > k \geq 1$. Now consider growing a branch based on this trunk: Add a cue combination that contains this subset of k cues and an extra cue m that is not any of the n cues.

Because the combinations in nested set A that are not in nested set B (i.e., those that contain any cue from Cue $k+1, k+2, \ldots,$ to Cue n) do not affect the weights of cues forming set B, the new set of combinations that consists of (a) set B and (b) the combination that contains Cue m is nested (i.e., cues from $k+1, k+2, \ldots,$ to n can be ignored with respect to the weights of the cues in this new set). Now, because the relation between the two branches that share B as their common trunk is symmetrical, the converse implication holds: Set A remains a nested set despite the addition of the new combination that contains cue m. Thus, Equation A20 applies to any cue in these two partially overlapping nested sets, and Equation A23 applies to any of these cues except the one in the smallest combination. The same argument applies to the growth of any branch anywhere on the tree if that branch does not contain any cue in other branches of the tree (see Figure 1 and its accompanying text for a visual characterization of a nested set that contains partially overlapping combinations).

My definition of *nesting* can therefore be generalized to cover partially overlapping designs. In a design with multiple cues, if there are no partially overlapping cue combinations unless for every pair of such combinations, all supersets of one combination share the same intersection with the other combination (i.e., they share a common trunk, and this trunk is the only thing they share), and this intersection occurs as a separate combination, then the design is nested. In other words, except for such partially overlapping combinations, every combination of stimuli in a nested design can be characterized as a proper superset of any combination that contains some but not all stimuli in it.

Asymptotic Weights for Designs for Which the R–W Model Does Not Compute Conditional Contrasts

My analysis of the conditions under which the R–W model computes conditional contrast shows that it does so when the cue combinations are nested. When the cue combinations are not nested, the strength of a cue is not necessarily equal to any of its (conditional or unconditional) contrasts. For example, consider the nonnested design mentioned earlier with the combinations *a*, *ab*, and *bc*. In this section, let the subscript denote the combination, so that π_a/k_a, π_{ab}/k_{ab}, and π_{bc}/k_{bc}, respectively, are the relative frequencies of the outcome j given combinations a, ab, and bc. First, consider the combinations a and ab, assuming that the bc combination does not exist. Because a is nested within ab, the preceding analysis of nested sets applies. Applying Equation A23 to this nested set, we obtain

$$V_b = \frac{\pi_{ab}}{k_{ab}} - \frac{\pi_a}{k_a}. \qquad \text{(A24)}$$

Now consider adding the bc combination. Because cue c appears only in this combination, Equation A17 applies. For the same reason, we know that adding this combination does not change the asymptotic weights of cues that appear in the other combinations. That is, it does not change V_a or V_b. Thus, by applying Equation A17 to the bc combination and substituting for V_b, we obtain

$$V_c = \frac{\pi_{bc}}{k_{bc}} - V_b = \frac{\pi_{bc}}{k_{bc}} - \left(\frac{\pi_{ab}}{k_{ab}} - \frac{\pi_a}{k_a}\right). \qquad \text{(A25)}$$

Because π_{bc}/k_{bc} is the frequency of the outcome given the presence of only b and c, Equation A25 shows that if and only if V_b is an estimate of the probability of outcome j occurring in the presence of b alone will V_c be equal to the contrast of Cue c conditional on the presence of Cue b, that is, $P(j|\bar{a}bc) - P(j|\bar{a}b\bar{c})$. In other words, unless Cue a can be ignored as a conditionalizing cue, as when it is believed not to cause the outcome (i.e., if $\pi_{ab}/k_{ab} = \pi_b/k_b$ and $\pi_a/k_a = 0$), V_c is not equal to the contrast for Cue c conditional on the sole presence of Cue b. In that special case in which a is not a plausible cause, the design is nested with respect to plausible causes.

Interpreting Equation A25 in terms of the causal powers of candidate causes a, b, and c, assuming that these are all of the cues in the focal set and that they produce effect j independently, we see that

$$\begin{aligned} V_c &= \frac{\pi_{bc}}{k_{bc}} - \left(\frac{\pi_{ab}}{k_{ab}} - \frac{\pi_a}{k_a}\right) \\ &= P(j|bc) - P(j|ab) + P(j|a) \\ &= p_b + p_c - p_b \cdot p_c - (p_a + p_b - p_a \cdot p_b) + p_a. \qquad \text{(A26)} \end{aligned}$$

Therefore,

$$V_c = p_c - p_b \cdot p_c + p_a \cdot p_b \qquad \text{(A27)}$$

Because the RHS of Equation A27 has both a positive and a negative term in addition to p_C, V_C is not interpretable as an estimate of the power of Cue c. In sum, when the cue combinations are not nested, the strength of a cue is not, in general, interpretable as a conditional contrast or an estimation of causal power.

Appendix B

When the R–W Model Computes Conditional Contrasts in an Induced Overshadowing Design

In an induced overshadowing design, denoting the two varying cues as a and b and the context as c, the design is c, ac, bc, abc. Let k denote the total number of trials with a certain cue combination, with its subscript denoting the combination (e.g., k_{abc}), and let π denote the number of trials for a cue combination on which the outcome occurs, with its subscript again denoting the combination (e.g., π_{abc}). First, consider the cue combinations c, bc, and abc in this design, ignoring the combination ac. These three combinations form a simple nested set. Therefore, according to Equation A23 (see Appendix A), the R–W model predicts that the strengths of cues a and b, V_a and V_b, respectively, are as follows:

$$V_a = \frac{\pi_{abc}}{k_{abc}} - \frac{\pi_{bc}}{k_{bc}} \qquad \text{(B1)}$$

and

$$V_b = \frac{\pi_{bc}}{k_{bc}} - \frac{\pi_c}{k_c}. \qquad \text{(B2)}$$

In addition, according to Equation A20,

$$V_c = \frac{\pi_c}{k_c}, \qquad \text{(B3)}$$

where V_c is the strength of c, the cue that occurs in isolation.

Recall that the R–W model (see Equation 15) predicts that the relative frequency with which the outcome occurs for a given cue combination is the sum of the strengths of the cues in that combination. Therefore, for combination ac, it predicts that

$$\frac{\pi_{ac}}{k_{ac}} = V_a + V_c. \qquad \text{(B4)}$$

This prediction would hold for the design (i.e., the R–W model would converge on a solution) if V_a and V_c in Equation B4 have the same respective values as in the nested set earlier (i.e., as in Equations B1 and B3). Thus, substituting for V_a and V_c in Equation B4 with their respective values in Equations B1 and B3, we obtain

$$\frac{\pi_{ac}}{k_{ac}} = \frac{\pi_{abc}}{k_{abc}} - \frac{\pi_{bc}}{k_{bc}} + \frac{\pi_c}{k_c}. \qquad \text{(B5)}$$

Rearranging this equation, we see that

$$\frac{\pi_{ac}}{k_{ac}} - \frac{\pi_c}{k_c} = \frac{\pi_{abc}}{k_{abc}} - \frac{\pi_{bc}}{k_{bc}}. \qquad \text{(B6)}$$

That is, the R–W model would converge on the same solution as in the nested set c, bc, and abc if the contrast for a conditional on the absence of b and the presence of c (i.e., $P(e|a\bar{b}c) - P(e|\bar{a}\bar{b}c)$), the LHS of Equation B6 is equal to its contrast conditional on the presence of b and c (i.e., $P(e|abc) - P(e|\bar{a}bc)$), the RHS of Equation B6).

The same argument applies when one considers the nested cue combinations c, ac, and abc in this design, ignoring the combination bc. More generally, then, the R–W model converges on the same solution as in a nested set within the induced overshadowing design if the contrast for a varying cue (e.g., a) conditional on the presence of the other varying cue and the context is equal to its contrast conditional on the absence of the other cue and the presence of the context.

Received January 7, 1994
Revision received July 2, 1996
Accepted July 2, 1996 ■