Journal of Experimental Psychology:
Learning, Memory, and Cognition
2003, Vol. 29, No. 6, 1119–1140

Copyright 2003 by the American Psychological Association, Inc.
0278-7393/03/$12.00 DOI: 10.1037/0278-7393.29.6.1119

From Covariation to Causation: A Test of the Assumption of Causal Power

Marc J. Buehner
Cardiff University

Patricia W. Cheng and Deborah Clifford
University of California, Los Angeles

How humans infer causation from covariation has been the subject of a vigorous debate, most recently between the computational causal power account (P. W. Cheng, 1997) and associative learning theorists (e.g., K. Lober & D. R. Shanks, 2000). Whereas most researchers in the subject area agree that causal power as computed by the power PC theory offers a normative account of the inductive process, Lober and Shanks, among others, have questioned the empirical validity of the theory. This article offers a full report and additional analyses of the original study featured in Lober and Shanks's critique (M. J. Buehner & P. W. Cheng, 1997) and reports tests of Lober and Shanks's and other explanations of the pattern of causal judgments. Deviations from normativity, including the outcome-density bias, were found to be misperceptions of the input or other artifacts of the experimental procedures rather than inherent to the process of causal induction.

> ". . . what is the nature of that evidence which assures us of any real existence and matter of fact, beyond the present testimony of our senses, or the records of our memory?"
>
> —David Hume (1777/1902, p. 26)

PURELY COVARIATIONAL MODELS

The Scottish philosopher David Hume (1739/1987) posed a fundamental problem concerning the acquisition of causal knowledge that has remained a puzzle to this day: Our sensory input does not contain explicit information about causal relations. Causal structure must be computed from information available to our senses—such as the presence and absence of candidate causes and effects—by some process of induction. Philosophers, learning theorists, social psychologists, and cognitive psychologists have debated how such information is used to infer causality. For candidate causes and effects represented by binary variables, a longstanding proposal (e.g., Jenkins & Ward, 1965) is that the evaluation of a relation between a candidate cause, *c,* and an effect in question, *e,* is based on

$$\Delta P = P(e|c) - P(e|\neg c), \tag{1}$$

where $P(e|c)$ is the probability of e given the presence of c, and $P(e|\neg c)$ is the probability of e given the absence of c. ΔP is a measure of the extent to which c and e covary and is often called the *contingency* or *contrast.* The conditional probabilities are estimated by the relevant relative frequencies. If ΔP is noticeably positive, c is a *generative* cause, and if it is noticeably negative, c is a *preventive* cause. If ΔP does not noticeably differ from zero (i.e., the candidate is noncontingent), c is independent of e and is noncausal.

A related attempt to explain the mental leap from covariation to causation likewise declares heritage to Hume (1739/1987): Causal judgment reflects "no more than the strength of the relevant association between the mental representations of cause and effect, with the principles governing such attributions being those of associative learning" (Shanks & Dickinson, 1987, p. 230). The most influential theory of associative learning has been the Rescorla-Wagner (Rescorla & Wagner, 1972) model (RWM). Although originally proposed to describe Pavlovian conditioning, a number of researchers have adopted the RWM to explain human causal reasoning (e.g., Shanks, 1985; Shanks & Dickinson, 1987; Wasserman, Elek, Chatlosh, & Baker, 1993). The core of the RWM is a mechanism of error correction or discrepancy reduction over repeated conditioning trials. Learning proceeds by changing the associative strength between a conditioned stimulus (CS; e.g., a flash of light) and an unconditioned stimulus (US; e.g., a shock). According to this model,

$$\Delta V_{CS} = \alpha_{CS} \cdot \beta_{US} \cdot (\lambda_{US} - \sum_{CS} V) \tag{2}$$

where ΔV_{CS} is the change in the strength of an association between a CS and a US after a given trial and α_{CS} and β_{US} are learning-rate parameters that respectively represent the salience of the CS (e.g.,

Marc J. Buehner, Department of Psychology, Cardiff University, Cardiff, Wales; Patricia W. Cheng and Deborah Clifford, Department of Psychology, University of California, Los Angeles.

Experiment 1 was part of a master's thesis submitted by Marc J. Buehner to the Department of Psychology, University of California, Los Angeles, under the supervision of Patricia W. Cheng, and was presented at the 19th Annual Meeting of the Cognitive Science Society, Stanford, California, 1997, and at the 39th Tagung experimentell arbeitender Psychologen in Berlin, Germany, 1997. Preparation of this article was supported by a Fulbright scholarship and a Friedrich-Ebert grant awarded to Marc J. Buehner and by National Science Foundation Grant 9729726, National Institutes of Health Grant MH64810, and a Guggenheim Fellowship awarded to Patricia W. Cheng.

Sean McAulliffe programmed the experimental materials for Experiment 1. We are grateful to Carla Chandler for suggesting simultaneous trials. We thank Merideth Gattis, David Shanks, Keith Holyoak, and one anonymous reviewer for valuable comments on a draft of the manuscript and John Pearce for his help in setting the parameters in his model.

Correspondence concerning this article should be addressed to Marc J. Buehner, School of Psychology, Cardiff University, P.O. Box 901, Cardiff, CF10 3YG, Wales, United Kingdom. E-mail: BuehnerM@Cardiff.ac.uk

the brightness of the light) and the US (e.g., the intensity of the shock). λ_{US} represents the actual outcome of the trial and is typically set to 1 if the US is present and to 0 otherwise. ΣV is the sum of the associative strengths of all CSs present and is, thus, the expected outcome. Learning is the reduction of discrepancy between the expected and the actual outcome. When this discrepancy ($\lambda_{US} - \Sigma V$) approximates zero, learning has reached asymptote: The outcome is fully explained by the CSs. Different assumptions about the learning parameter β lead to two different versions of the RWM. The restricted RWM operates with ΔP held constant across trials on which the US is respectively present and absent ($\beta_{US} = \beta_{\overline{US}}$). The unrestricted RWM allows values of β to differ between these two types of trials; usually the presence of the US is assumed to be more salient than its absence, yielding $\beta_{US} > \beta_{\overline{US}}$ (cf., e.g., Miller, Barnet, & Grahame, 1995).[1] This ordering of parameter values is required for explaining the relative validity of generative candidate causes (Shanks, 1991; cf. Rescorla & Wagner, 1972; Wagner, Logan, Haberlandt, & Price, 1968).

Explaining human causal induction with the RWM reduces causal reasoning to associative learning: The candidate cause *c* is mapped onto the CS, the effect *e* onto the US, and the causal strength of *c* is mapped onto *c*'s associative strength. When there is only one varying candidate cause in a context, the restricted RWM asymptotically computes ΔP as the measure of associative strength (Chapman & Robbins, 1990). The unrestricted RWM, however, does not compute ΔP: With $\beta_{US} > \beta_{\overline{US}}$, it predicts that for any fixed positive or negative ΔP, the (absolute) magnitudes of the judged causal strengths will be smaller as $P(e|\neg c)$, the base rate of *e*, increases (see Wasserman et al., 1993). But if $\beta_{US} < \beta_{\overline{US}}$, the RWM predicts the opposite trend: The (absolute) magnitudes of the judged causal strengths should be larger as $P(e|\neg c)$ increases for any fixed positive or negative ΔP. Regardless of assumptions about β as long as the same ordering of its values is assumed, the RWM predicts $P(e|\neg c)$ to influence the absolute causal strengths for candidates with equal positive ΔP in the same direction as those with equal negative ΔP. These predictions are graphed in Figure 1, which displays detailed predictions for Experiment 1 (discussed in the *Method* section of that experiment).

The German critical philosopher Immanuel Kant (1781/1965) and many others have noted that covariation or regular succession does not necessarily imply causation (a point typically emphasized in any course on experimental design). However, both the contingency model and the RWM disregard Kant's arguments and treat covariation as a direct measure of causation. These models, being purely covariational, do not represent causal strength as a variable that is separate from covariation. As a result, such models cannot acquire a state of knowledge in which causal strength is unknown (i.e., has no value) when covariation has a definite value (Wu & Cheng, 1999). Consider a concrete illustration of this problem. Suppose a researcher wants to evaluate the preventive strength of a new headache-relieving drug. In a study involving 16 participants, 8 receive treatment with the drug (candidate present), and 8 receive a placebo (candidate absent). Now suppose neither the 8 participants who received treatment nor the 8 control participants report headaches. ΔP equals 0, yet the researcher would not infer that the drug is ineffective (i.e., noncausal). Because headaches did not occur even in the control group, how then could a preventive candidate show that it has prevented them in the treatment group? This situation is the preventive analog of a ceiling effect in evaluating generative causes. In both circumstances, one simply cannot draw any causal inferences, even though the covariation has a definite value, 0.

THE POWER PC THEORY

To account for the distinction between covariation and causation, Cheng (1997) proposed the power PC theory. According to this theory, the causal reasoner's goal is to induce the unobservable *causal power* of a candidate cause in the distal world from observable events represented in the proximal stimulus. A candidate's causal power is the probability with which it influences the effect in question, either producing or preventing it. For the case in which ΔP is nonnegative, the theory shows that when (and only when) causes alternative to the candidate cause *c* both occur and influence *e* independently of *c*, the generative power of *c* to produce *e* is

$$q = \frac{\Delta P}{1 - P(e|\neg c)}. \tag{3}$$

For a nonpositive ΔP, the same set of assumptions except that *c* may now potentially prevent instead of produce *e* yields

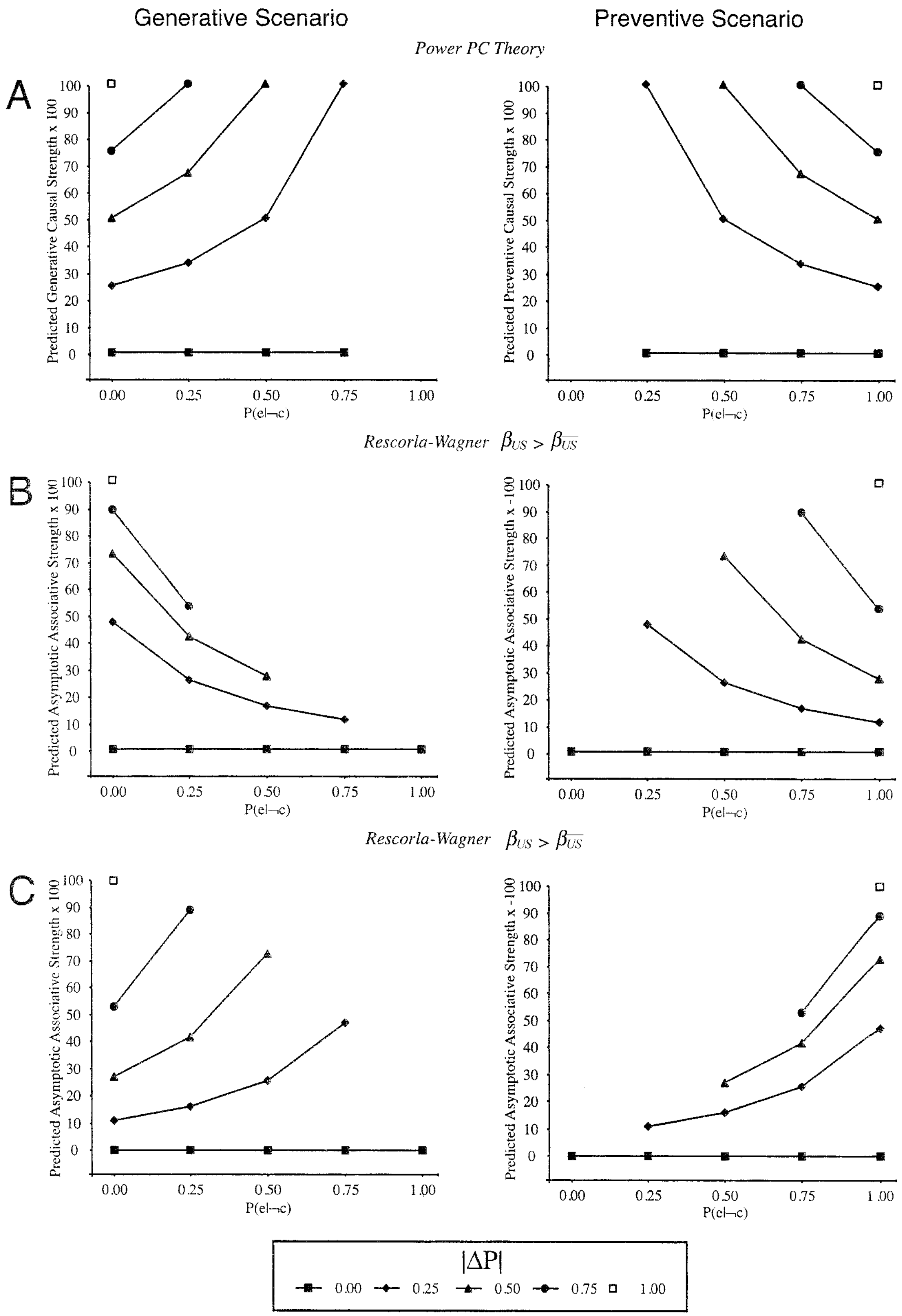
$$p = \frac{-\Delta P}{P(e|\neg c)} \tag{4}$$

as the causal power of *c* to prevent *e*. Note that the use of one rather than the other of these two equations is determined by the sign of ΔP, which can be computed from the observed input, rather than by any parameter settings. (If $\Delta P = 0$, both equations would be applicable.)

Under conditions in which Equations 3 and 4 apply, causal power depends not only on ΔP but also on $P(e|\neg c)$. It follows from these equations that candidate causes with equal levels of nonzero ΔP but different values of $P(e|\neg c)$ should yield different causal judgments. (Conversely, it is possible for candidates with different nonzero values of ΔP to have the same value of causal power,

[1] Lober and Shanks (2000) coined the terms *restricted* versus *unrestricted* Rescorla-Wagner model.

Figure 1 (*opposite*). A: Predictions of the power PC theory. B: Predictions of the Rescorla-Wagner (1972) model with $\beta_{US} > \beta_{\overline{US}}$. C: Predictions of the Rescorla-Wagner model with $\beta_{US} < \beta_{\overline{US}}$ for the generative (left) and preventive (right) components of Experiment 1. Lines connect conditions that share the same level of ΔP. Note that noncontingent conditions with undefined generative and preventive powers—$\Delta P = P(e|c) - P(e|\neg c) = 1.00 - 1.00$ and $0.00 - 0.00$, respectively—are excluded from the prediction plots for the power PC theory. The respective values for β_{US} and $\beta_{\overline{US}}$ are .8 and .3 in Panel B and .3 and .8 in Panel C.


Generative Scenario
Preventive Scenario
Power PC Theory
A
Predicted Generative Causal Strength x 100
Predicted Preventive Causal Strength x 100
0 10 20 30 40 50 60 70 80 90 100
0.00 0.25 0.50 0.75 1.00
P(e|¬c)
Rescorla-Wagner $\beta_{US} > \beta_{\overline{US}}$
B
Predicted Asymptotic Associative Strength x 100
Predicted Asymptotic Associative Strength x -100
Rescorla-Wagner $\beta_{US} > \beta_{\overline{US}}$
C
Predicted Asymptotic Associative Strength x 100
Predicted Asymptotic Associative Strength x -100
|ΔP|
0.00
0.25
0.50
0.75
1.00

hence predicting equal causal judgments.) In particular, when evaluating generative causal power, candidates with the same positive ΔP should be judged to have increasingly large generative power as $P(e|\neg c)$ increases toward, but does not equal, 1. When $P(e|\neg c) = 1$, q has an undefined value according to Equation 3: A reasoner cannot draw any conclusion about the causal strength of c generating e if e occurs all the time regardless of c (consistent with the avoidance of statistical inference in the case of a ceiling effect). In contrast, candidate causes with equal negative values of ΔP should be judged to have increasingly small preventive power as $P(e|\neg c)$ increases. When $P(e|\neg c) = 0$, p has an undefined value according to Equation 4, as in the headache-drug example where preventive power could not be assessed. Finally, Equations 3 and 4 both predict that when $\Delta P = 0$, the power of c should remain at 0 and be uninfluenced by $P(e|\neg c)$ as long as the denominator in the relevant equation is not 0. These predictions are assumed to apply only as ordinal descriptions.

To illustrate these predictions, let us return to the researcher who wants to evaluate the effectiveness of new headache-relieving drugs: Again, 8 participants receive treatment and 8 a placebo, assuming that all alternative causes of headaches are constant across the two groups. Now, all 8 participants in the control group have headaches, whereas only 6 of the 8 participants who received the drug have headaches, $\Delta P = P(e|c) - P(e|\neg c) = .75 - 1.00 = -.25$. The researcher reasons that if not for the drug, all 8 participants in the drug group would have had headaches, just as in the control group. The drug therefore has a small preventive power, preventing headaches with a probability of .25. In yet another study, 4 of the 8 participants in the control group report headaches and 2 of the 8 participants in the treatment group report headaches, $\Delta P = P(e|c) - P(e|\neg c) = .25 - .50 = -.25$. Alternative causes would have produced headaches in 4 of the 8 participants in the drug group, just as in the control group. The drug therefore prevents headaches in 2 of these 4 participants, yielding a probability of .50. Thus, although $\Delta P = -.25$ here as in the preceding study, the researcher will attribute a higher preventive power to the latter candidate. Equation 4 formalizes this intuition. Analogous intuitions about the opposite influence of $P(e|\neg c)$ on the generative power of candidate causes with the same positive ΔP are captured by Equation 3.

ALTERNATIVE PURELY COVARIATIONAL APPROACHES

Recall that, in contrast to the power PC theory, the unrestricted RWM predicts the same influence of $P(e|\neg c)$ for candidates with positive or negative ΔPs, as long as the ordering of β values remains unchanged (e.g., $\beta_{US} > \beta_{\overline{US}}$ throughout). But, there are a number of purely covariational models that can capture the interactive influence of base rate and the sign of ΔP on judged causal strength. Unlike the power PC theory, however, this increase in explanatory power is achieved by introducing additional parameters. For example, in the weighted ΔP model (e.g., Wasserman et al., 1993; also see Anderson & Sheu, 1995), the two conditional probabilities in Equation 1 may assume different weights, denoted by w_1 and w_2 below:

$$\Delta P = w_1 P(e|c) - w_2 P(e|\neg c). \tag{5}$$

According to the linear combination models (Schustack & Sternberg, 1981), causal strength S is a linear function of the frequencies of the four possible types of events involving c and e, with w_a and w_d in Equation 6 typically given a positive value and w_b and w_c typically given a negative value, $w_\text{a} > w_\text{d}$, $|w_\text{b}| > |w_\text{c}|$, and i as the intercept:

$$S = i + w_a \cdot P(e|c) + w_b \cdot P(\neg e|c) + w_c \cdot P(e|\neg c) + w_d \cdot P(\neg e|\neg c). \tag{6}$$

Another model of this type is Pearce's (1987) associative model of stimulus generalization. The main difference between Pearce's model and the RWM is that the former, unlike the latter, does not accrue associative strengths of individual cues. In Pearce's model, any combination of cues (including the combination of only one cue with a constant learning context) is represented as a compound cue. The associative strength of a single cue alone has to be computed by multiplying the strength of the compound in question with a parameter representing the similarity of the single cue to the compound. This extra parameter, when attached to the various additional representations of the strengths of the compound cues, allows additional degrees of freedom to provide a better fit of associative learning theory to data from causal induction experiments. Some researchers have chosen to take this path (e.g., Baker, Vallee Tourangeau, & Murphy, 2000; Perales & Shanks, in press; Vallee Tourangeau, Murphy, Drew, & Baker, 1998).

All three models are indeed able to qualitatively account for a wider range of findings compared with purely covariational models with fewer parameters. Provided that the parameters are set the right way, they can predict an interactive influence of the sign of ΔP and the base rate of e on causal ratings for candidates with identical levels of nonzero ΔP (see also Perales & Shanks, in press). Note that all three models, unlike the traditional ΔP theory, the power PC theory, and both versions of RWM, predict that differences in the base rate of e should also affect estimated causal strength in noncontingent candidates, just as in contingent candidates with identical causal power.

EXPERIMENTAL TESTS OF THE TWO APPROACHES

Experiment 1

The goal of the first experiment was to test the interactive influence of base rate and the sign of ΔP on judged causal strength on contingent, but not on noncontingent, candidates, as predicted by a computational causal power approach (the power PC theory), to distinguish it from purely covariational accounts. We tested sets of candidates with positive, negative, and zero ΔPs. A comprehensive test of these two approaches would include combinations of many levels of $P(e|c)$ and $P(e|\neg c)$. Wasserman et al. (1993) reported the most exhaustive study of this nature to date but used an effect that occurred at a rate per unit of continuous time rather than in a proportion of discrete entities. The power PC theory is a probabilistic theory and hence does not apply to the type of outcome variable in Wasserman et al. We therefore modified their design in which five levels of $P(e|c)$ and of $P(e|\neg c)$ are independently combined within-subject by using a binary effect that occurred in discrete entities instead of in continuous time. Because such entities required a longer presentation time, we presented two separate groups of participants with conditions involving nonnegative ΔPs and those involving nonpositive ΔPs to keep the number of conditions manageable for a participant.

Method

Participants

Fifty-seven (preventive component) and 52 (generative component) undergraduate psychology students from the University of California, Los Angeles, participated to partially fulfill a course requirement.

Design and Procedure

Participants in the preventive component (involving nonpositive ΔPs) were given a cover story asking them to pretend they were virologists testing several new vaccines against viruses. Each participant worked on 1 practice condition and 15 experimental conditions, with each condition consisting of 16 laboratory records (i.e., learning trials). Each record provided information about whether one particular rat was vaccinated prior to virus exposure and whether this rat developed the disease related to the virus. The generative component (involving nonnegative ΔPs) adopted the same design and procedure, except that participants were asked to imagine that they were microbiologists studying how ray exposure influences the mutation of viruses; laboratory records informed participants whether one particular petri dish with viruses was exposed to certain rays and whether the viruses in this dish mutated. The 15 experimental conditions in both components of the experiment represented 15 independent studies on different viruses and vaccines or rays and viruses. We used 16 fictitious viruses, vaccinations, and rays.

For the evaluation of preventive causal power, five levels (1.00, .75, .50, .25, and .00) of the conditional probabilities $P(e|c)$ and $P(e|\neg c)$ were combined to yield five levels of nonpositive ΔPs: −1.00, −.75, −.50, −.25, and .00. These combinations rendered a total of 15 conditions (see Table 1), which were presented in random order. For the evaluation of generative causal power, the design was exactly symmetrical, producing nonnegative ΔPs with values .00, .25, .50, .75, and 1.00 (see Table 2).

The 16 laboratory records for the evaluation of each candidate consisted of 8 for which *c* was present (the rat was vaccinated or the petri dish was exposed to rays) and 8 for which *c* was absent (no vaccination given or no ray exposure). These records were presented in random order sequentially on a computer screen. For each record, participants had to confirm the information provided by clicking appropriate checkboxes on the computer screen ("received vaccination? yes/no" and "broke out with virus disease? yes/no" in the preventive part and "was exposed to rays? yes/no" and "mutated? yes/no", in the generative part).

After having studied the 16 records in a condition, participants evaluated the causal power of the studied candidate. Participants in the preventive part were asked to judge how strongly each vaccine prevented the disease related to the virus in question by giving a rating on a scale from 0 (*the vaccine does not prevent the disease at all*) to 100 (*the vaccine prevents the disease every time*). Analogously, participants in the generative part of the study were asked to rate how strongly they thought the particular rays cause mutation on a scale from 0 (*the ray does not cause mutation at all*) to 100 (*the rays cause mutation every time*).[2] To keep the judgments simple, the task did not provide the option of answering "I don't know". Following their answer, participants' confidence in their judgment as well as their subjective probability ratings of $P(e|c)$ and $P(e|\neg c)$ were collected.[3] The questions were always presented sequentially in the order just described. A graphical scale with corresponding labels at the endpoints and numerical markings at regular intervals was presented along with the text for the causal and confidence ratings. All participants were tested individually, and the experiment took approximately 45 min.

Table 1
Design and Results of Experiment 1: Preventive Component

				Causal ratings	
$P(e\|c)$	$P(e\|\sim c)$	ΔP	Causal power	*M*	*SD*
0.00	1.00	−1.00	1.00	93.58	12.87
0.00	0.75	−0.75	1.00	84.75	19.72
0.25	1.00	−0.75	0.75	71.54	17.26
0.00	0.50	−0.50	1.00	79.28	20.54
0.25	0.75	−0.50	0.67	64.61	19.49
0.50	1.00	−0.50	0.50	45.95	21.87
0.00	0.25	−0.25	1.00	72.00	29.62
0.25	0.50	−0.25	0.50	58.56	20.02
0.50	0.75	−0.25	0.33	42.86	21.83
0.75	1.00	−0.25	0.25	21.09	15.8
0.00	0.00	0.00	Undefined	45.12	41.05
0.25	0.25	0.00	0.00	47.79	28.70
0.50	0.50	0.00	0.00	33.89	23.06
0.75	0.75	0.00	0.00	22.86	21.86
1.00	1.00	0.00	0.00	8.68	20.39

Note. $N = 57$. Because preventive and generative powers are probabilities (Cheng, 1997), we refer to both with positive numbers.

Predictions

Figure 1 displays the pattern predicted by the power PC theory and by two versions of the unrestricted RWM. We only discuss the asymptotic predictions of the RWM. Predictions for the power PC theory result from simply inserting the conditional probabilities from each of the 15 generative and preventive conditions into Equations 3 and 4; which equation applies depends on the sign of ΔP, which as we mentioned is determined by the observable covariation alone. Asymptotic[4] predictions for the RWM cannot be made without making assumptions about the values of the

[2] One reviewer pointed out that this labeling of the extreme ends of the scale might encourage a deterministic as opposed to a probabilistic notion of causality, contrary to the scope of the theories considered here. We fully agree that the labels on the extreme ends do refer to deterministic relations, but this is consistent with a probabilistic account: 0% probability is equivalent to "not at all" and 100% probability to "every time", and intermediate values reflect intermediate probabilities.

[3] For the sake of brevity, we report the causal ratings. Confidence ratings displayed similar qualitative patterns as found in causal ratings and some aspects of the subjective probability ratings are considered in the discussion.

[4] The error-correcting algorithm of the RWM (and all other models of associative learning) actually never reaches asymptote, unless a cue is deterministic. Rather, associative strengths will always oscillate from trial to trial around a hypothetical asymptote; the residual error range is directly proportional to the values of the learning parameters. The standard way to derive a theoretical asymptote for associative models is to set the expected value of the ΔVs to zero and thus compute the "asymptotic" associative strengths (e.g., Wasserman et al., 1993). Strictly speaking it would be more accurate to refer to an equilibrium instead of an asymptote (see Danks, in press, who also offers a more general analysis of how to derive equilibria of the RWM in situations with more than one cue).

Table 2
Design and Results of Experiment 1: Generative Component

				Causal ratings	
$P(e\|c)$	$P(e\|\sim c)$	ΔP	Causal power	*M*	*SD*
1.00	0.00	1.00	1.00	88.54	20.05
1.00	0.25	0.75	1.00	76.54	23.86
0.75	0.00	0.75	0.75	69.02	22.21
1.00	0.50	0.50	1.00	70.85	23.19
0.75	0.25	0.50	0.67	54.02	23.77
0.50	0.00	0.50	0.50	57.00	22.73
1.00	0.75	0.25	1.00	57.62	32.35
0.75	0.50	0.25	0.50	47.06	26.49
0.50	0.25	0.25	0.33	45.71	22.53
0.25	0.00	0.25	0.25	33.69	26.67
1.00	1.00	0.00	Undefined	41.10	36.85
0.75	0.75	0.00	0.00	42.62	27.06
0.50	0.50	0.00	0.00	36.60	24.91
0.25	0.25	0.00	0.00	28.62	21.45
0.00	0.00	0.00	0.00	9.46	20.03

Note. $N = 52$.

learning parameter β.[5] For the atypical and problematic parameter ordering of $\beta_{US} < \beta_{\overline{US}}$, we used the values adopted by Lober and Shanks (2000): $\beta_{US} = .3$ and $\beta_{\overline{US}} = .8$. For the typical assumption of $\beta_{US} > \beta_{\overline{US}}$, we chose the reverse ordering. We derived predictions for the RWM using the formula offered by Wasserman et al. (1993):

$$V_{asymp} = \frac{\beta_{US}a}{\beta_{US}a + \beta_{\overline{US}}b} - \frac{\beta_{US}c}{\beta_{US}c + \beta_{\overline{US}}d}, \qquad (7)$$

where *a, b, c,* and *d* refer to the four cells of the contingency table displayed in Figure 2.

The lines in Figure 1 connect predicted ratings for the same level of ΔP. As the figure clearly illustrates, the power PC theory predicts positive slopes for generative candidates with identical levels of positive ΔP as the base rate of $eP(e|\neg c)$ increases and negative slopes for preventive candidates with identical levels of negative ΔP as the base rate of *e* increases. Figure 1 also shows that, depending on which ordering of parameters is used, the RWM predicts either an increase or a decrease for candidates of equal nonzero ΔP as the base rate of *e* increases but never an interaction between the base rate and the sign of ΔP. Both the power PC theory and the RWM predict noncontingent candidates to elicit noncausal ratings, regardless of variations in the base rate of *e*.

The points on the top edge of the two graphs displaying the predictions of the power PC theory (the two upper graphs) all share a predicted causal strength of 1.00. For the generative graph, these points all have a value of $P(e|c) = 1.00$; for the preventive graph, the corresponding points all share $P(e|c) = 0.00$. These predictions follow from Equations 3 and 4: Causal power is 1.00 when *e* always occurs or never occurs in the presence of *c*, regardless of variations in the base rate of *e*.

Predictions of the classic contingency model and the restricted RWM are not plotted in Figure 1; both of these models predict that conditions with identical ΔP should receive identical causal ratings, regardless of the base rate of *e*. Predictions for Pearce's (1987) model and the linear combination and weighted ΔP model are not plotted to save space but are considered in the *Discussion* section.

Results

Figure 2 displays participants' mean ratings of the preventive and generative causal power of the candidate causes. On the abscissa are the five levels of $P(e|\neg c)$. The lines connect mean ratings for the same level of ΔP. Values for $P(e|c)$ follow from the corresponding combination of ΔP and $P(e|\neg c)$. The means and standard deviations are listed in Tables 1 and 2.

Candidates With the Same ΔP but Different Causal Powers

A visual examination of Figure 2 reveals a substantial influence of $P(e|\neg c)$ on causal judgments for candidates with the same ΔP. As predicted by the power PC theory, this influence occurred in opposite directions for the preventive and generative causal ratings. We used one-factor repeated measures analyses of variance (ANOVAs) to examine trends that were due to $P(e|\neg c)$ in conditions with equal levels of ΔP, using an alpha level of .05 in all statistical tests. Linear negative trends were highly reliable for each level of negative ΔP: $t(56) = 4.14$ for $\Delta P = -.75$; $t(112) = 9.48$ for $\Delta P = -.50$; and $t(168) = 13.34$ for $\Delta P = -.25$, and no quadratic or cubic trends were found. For the generative part of the study, causal judgments for equal levels of ΔP generally increased as $P(e|\neg c)$ increased. For $\Delta P = .75$, however, the positive linear trend fell short of significance, $t(51) = 1.99$, $p = .52$. Linear positive trends were reliable for $\Delta P = .50$, $t(102) = 3.72$ (in addition to a cubic trend, $t[102] = 3.26$), and for $\Delta P = .25$, $t(153) = 3.91$.

This pattern of negative trends for candidates with the same negative ΔP and positive trends for candidates with the same positive ΔP clearly contradicts the traditional contingency model and the restricted RWM, as both postulate ΔP as the sole determinant of causal judgments and thus predict flat lines. The results also contradict the unrestricted version of the RWM with a consistent ordering of parameter settings, which could explain trends for equal levels of ΔP but merely in a single direction across preventive and generative scenarios.

Noncontingent Candidates

Participants' evaluations of noncontingent candidate causes substantially deviated from zero. In particular, there was a base-rate influence on causal ratings for noncontingent candidates in opposite directions in the generative and preventive conditions for the exact same set of five noncontingent conditions. In the preventive study, the 4 noncontingent candidates with defined causal powers (see Table 1) produced a negative linear trend, $t(168) = 9.57$, and the 4 corresponding noncontingent candidates in the generative study (see Table 2) produced a positive linear trend, $t(153) = 7.59$, and a cubic trend, $t(153) = 3.52$. This pattern of results, often called the outcome density bias, replicates earlier observations (e.g., Shanks, 1985, 1987) but is at variance with the power PC theory and asymptotic predictions of all variants of the RWM.

Candidates With the Same Causal Power but Different ΔPs

Our experimental design also allows comparisons between conditions with varying ΔPs for which causal power is predicted to be

[5] The learning parameter α is irrelevant for this analysis, because it only determines how soon the model reaches "asymptote" (see also Danks, in press).

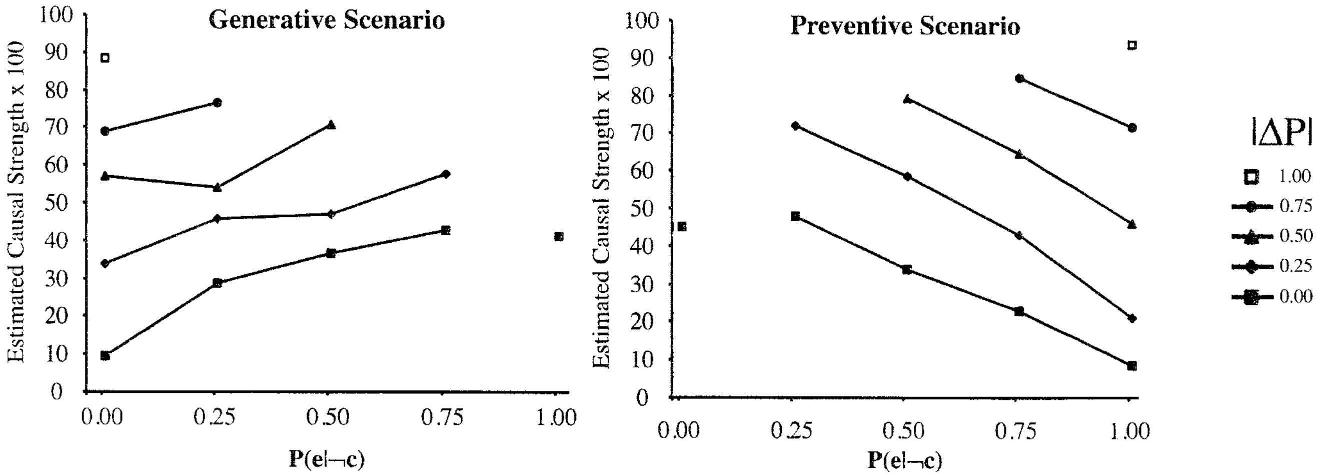


Figure 2. Experiment 1: Means of observed estimates of preventive and generative causal powers as a function of $P(e|\neg c)$. Lines connect conditions that share the same level of ΔP but have different causal powers. The unconnected single points with zero ΔP have undefined causal power.

constant. Among the preventive conditions, the 0/8–2/8, 0/8–4/8, 0/8–6/8, and the 0/8–8/8 conditions all yielded $p = 1$. In addition, the 2/8–4/8 and 4/8–8/8 conditions both yielded $p = .5$. Analogously, there were four generative conditions with $q = 1$ (8/8–0/8, 8/8–2/8, 8/8–4/8, and 8/8–6/8) and two with $q = .5$ (4/8–0/8 and 6/8–4/8). We analyzed conditions with equal causal power with repeated measures ANOVAs, with ΔP as the within-subject factor. Linear trends were reliable in both the preventive conditions, $t(168) = 5.82$, and the generative conditions, $t(153) = 6.16$, with causal powers of 1. Conditions with power equal to 0.5 differed significantly, $t(56) = 3.54$, for the preventive candidates, and $t(51) = 2.26$, for the generative candidates. These ΔP biases across conditions with identical causal powers are clearly problematic for the power PC theory.

Discussion

Refutation of the Classic Contingency Model, and Both the Restricted and Unrestricted RWM, as Adequate Accounts of Human Causal Induction

Experiment 1 demonstrated a substantial influence of the base rate of *e* in opposite directions for preventive and generative causal ratings when ΔP was kept constant. This pattern of results, subsequently replicated by Perales and Shanks (in press), effectively refutes the restricted RWM and contingency account as adequate accounts of human causal induction. The unrestricted RWM likewise cannot explain the interaction between causal scenario (preventive vs. generative) and the direction of the base-rate influence unless one assumes reversed settings of the values of Learning Parameter β across scenarios.

Lober and Shanks (2000) argued that different experimental contexts and cover stories justify such a reversal of parameter values, allowing the unrestricted RWM to explain the results. This argument is not only post hoc but also requires that the RWM sets $\beta_{US} < \beta_{\overline{US}}$, atypically, for generative causal relations, leaving the model unable to explain other well-established causal reasoning phenomena involving generative causal relations, such as relative validity (Shanks, 1991; cf. Rescorla & Wagner, 1972; Wagner, Logan, Haberlandt, & Price, 1968). Nonetheless, a clearer and stronger test of the power PC theory against the RWM would indeed include both preventive and generative ratings in a single experimental setting, sharing an identical cover story.

Problematic Results for the Power PC Theory

Experiment 1 also demonstrated two findings that appear to contradict the predictions of the power PC theory. First, there was a substantial influence of the base rate of *e* on the causal ratings of noncontingent candidates. This result contradicts not only the power PC theory but also both versions of RWM. Second, there was a substantial influence of ΔP on causal ratings of candidates with the same causal power, as the unrestricted RWM with inconsistent parameter settings can explain. Both of these deviations from the predictions of the power PC are consistent with Pearce's (1987) model, the weighted ΔP model (Anderson & Sheu, 1995), and linear combination models (Schustack & Sternberg, 1981) with their typical ordering of parameter weights. We performed linear regressions on the data from Experiment 1 to find the best fit for the latter two models. Because our materials presented the same number of trials when *c* was present as when it was absent, $P(e|c)$ and $P(\neg e|c)$, and $P(e|\neg c)$ and $P(\neg e|\neg c)$, respectively, are perfectly (negatively) correlated; the regression for the linear combination model therefore produced nonzero weights for only one member of each of these pairs of variables. It follows that for our materials, the linear combination model and the weighted ΔP model are the same except that the latter has an intercept restricted to 0. Figure 3 displays the best fit for the data from Experiment 1 for the linear combination model, the model with the better fit between these two. We have calculated separate regressions for the generative and preventive scenarios, and the respective weights obtained for the former are as follows: $w_1 = 8.89$, and $w_3 = -5.14$, whereas for the latter they are $w_1 = -10.89$, and $w_3 =$

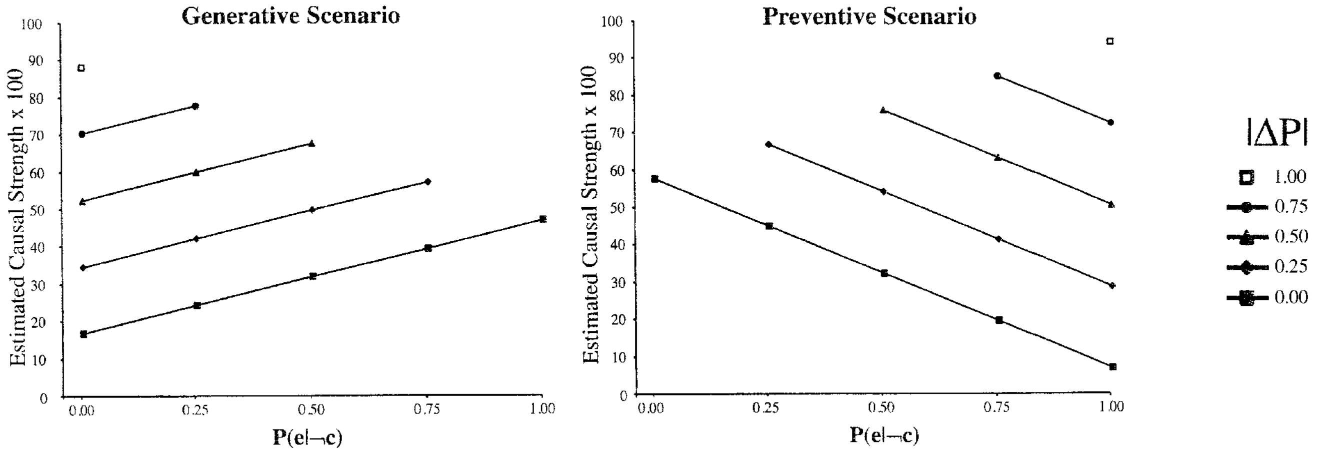


Figure 3. Best fit of the linear combination model to the causal ratings observed in Experiment 1 as a function of $P(e|\neg c)$. The weights were as follows: $w_a = -10.899$, $w_c = 4.557$, and $i = 57.597$, for the preventive component and $w_a = 8.894$, $w_c = -5.138$, and $i = 16.831$, for the generative component.

4.56. Adopting separate sets of parameters for generative and preventive scenarios was necessary to account for the opposite influence of base rate on noncontingent conditions across the two scenarios.

Do Deviations From Causal Power Reflect Fundamental Properties of the Reasoning Process, or Ambiguities and Extraneous Influences in the Experimental Method?

The linear combination model, the weighted ΔP model, and Pearce's (1987) model all predict that the deviations from causal power discussed above follow necessarily from the covariational input. Alternatively, these deviations may reflect (a) ambiguities in the stimulus materials used and (b) misrepresentation of the input that was due to memory limitations in this and similar experiments (e.g., Lober & Shanks, 2000).

Ambiguity of the causal question. One possible explanation for these deviations is that the rating scale that was used in Experiment 1 and other studies was ambiguous. Buehner and Cheng (1997) explained that the ambiguity might have led subjects to conflate reliability with causal strength. Tenenbaum and Griffiths (2001) proposed in a similar vein that subjects might have rated how confident they were that there was a causal relation (see their Bayesian causal support model). We consider another possible ambiguity here. The question might have given rise to two interpretations, one of which can produce causal ratings that map onto ΔP. How strongly a candidate cause produces (or prevents) an effect is ambiguous with respect to the context in which this question applies: (a) the current learning context or (b) a counterfactual context in which there are no alternative causes of like kind. Under the first interpretation, ΔP would be the rational answer, even if the participants did infer causal power. The question might have been interpreted as, "What difference does the candidate cause make in the current learning context, in which alternative causes already produce *e* in a certain proportion of the entities?" Consider a participant who infers that causal power is 2/3 in the 6/8–2/8 condition. Under the first interpretation, he or she would answer that the candidate would increase the occurrence of *e* by 50%: percentage of entities in the learning context in which alternative causes do not already produce *e* × causal power = (100% − 25%) × 2/3 = 50%, which of course corresponds to a ΔP of .50. Under the second interpretation, however, the question would be interpreted as, "What difference does the candidate cause make when alternative causes never produce *e?*" The answer to such an interpretation of the question (it would affect 2/3 of the entities in the counterfactual context) would correspond to causal power.[6]

Given that the rating scale we used in Experiment 1 allows both interpretations just discussed, it is plausible that some of the participants parsed the rating instructions the first way, yielding ratings directly related to ΔP, whereas others adopted the other valid reading of the instructions and based their answers on causal power. An approximately even split of participants across these two interpretations would in fact produce the data pattern we found for contingent candidates.

To test whether participants in Experiment 1 were indeed split into two such subgroups, we performed K-means cluster analyses on both datasets. We included all 15 conditions in the respective analyses and formed two clusters in each. The results of the cluster analysis support our hypothesis: Participants were indeed best split into those who primarily based their ratings on causal power (Subgroup A: $n = 33$ for the generative part; $n = 23$ for the preventive part) and on ΔP (Subgroup B: $n = 19$ for the generative part; $n = 34$ for the preventive part). Figure 4 displays the mean observed estimates of generative and preventive causal strength, respectively, in Subgroups A and B. Note that the lines connecting conditions with identical ΔPs have considerably steeper slopes in both graphs for Subgroup A compared with those for Subgroup B. In particular, data points on the top edge of each graph represent-

[6] Causal ratings directly corresponding to ΔP would of course also be obtained if participants merely indicated how much a cause increases or decreases the probability of the effect, as the ΔP model states. However, if participants who give a ΔP response simply obey this model, then clarifying the question should make no difference (see Experiment 2).

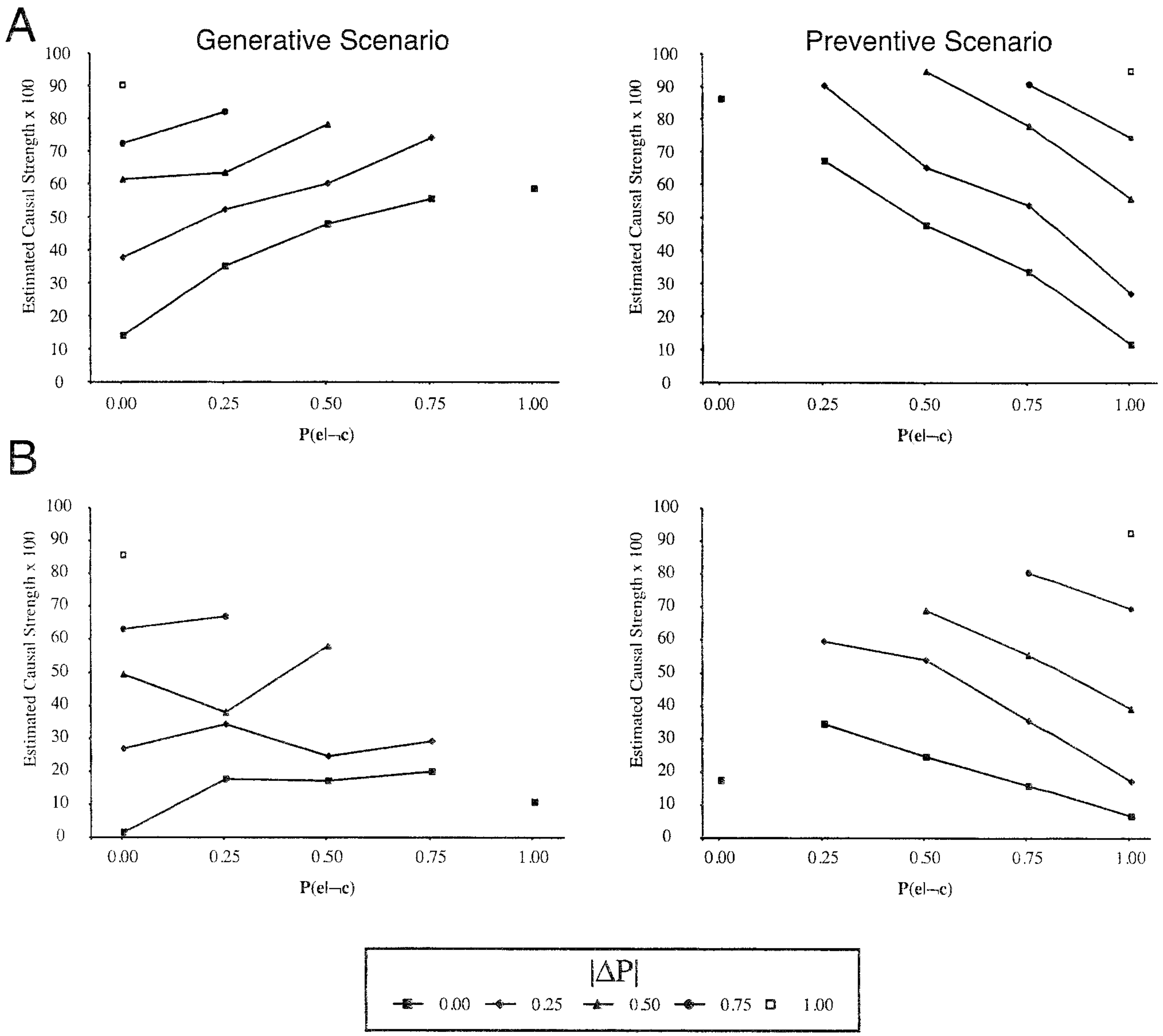


Figure 4. Experiment 1: Means of observed estimates of preventive and generative causal powers as a function of $P(e|\neg c)$ in participants whose ratings of causal candidates were driven primarily by causal power (A) and ΔP (B).

ing conditions with causal power of 1.0 are notably less influenced by variations in ΔP for Subgroup A than Subgroup B. Because this strategy split can be due to the ambiguity in the rating instructions used in Experiment 1 and other studies, permitting two valid interpretations of the context to which the rating refers, subsequent work should attempt to eliminate this ambiguity. Because the causal power interpretation is the only one that would differentiate between the ΔP model and the power PC theory, the causal question should unambiguously prompt for that interpretation. Among other improvements, Experiment 2 was designed to serve that purpose.

Misrepresentations due to memory limitations: Explaining the outcome-density bias. The ambiguity of context cannot account for the outcome-density effect for the noncontingent conditions, however. Regardless of whether one assumes the current learning context or a counterfactual context, if the candidate is judged to have no causal power, then introducing it to either context should make no difference.

The fact that the base rate of *e* interacted with the context (preventive vs. generative) to influence causal ratings of both noncontingent and contingent candidates suggests an explanation for these deviations along the lines of the computational causal power account. If participants misperceived that the relations in the noncontingent conditions were contingent, even if only by a small amount, as is plausible in a sequential trials design, the power PC theory of course would predict the same influence of base rate for these misperceived contingencies as it does for correctly perceived contingent candidates (i.e., generative ratings should increase and preventive ratings should decrease, as base rate increases).

Support for this interpretation can be found by analyzing participants' subjective estimates of the conditional probabilities. We calculated participants' subjective impressions of ΔP by subtracting their estimates of $P(e|\neg c)$ and $P(e|c)$ from each other. For the five noncontingent conditions in the generative component, the overall obtained mean was 3.75, which is significantly different from 0, $t(259) = 2.39$, $p < .02$; for the five noncontingent conditions in the preventive component, the overall obtained mean was -3.08, again significantly different from 0, $t(284) = 2.39$, $p < .02$. Thus, it is plausible that the observed outcome-density effect was due to a misperception that the contingencies were positive or negative, when in fact they were zero. To differentiate between the competing accounts of causal induction, we reduced memory demands in Experiment 2. In addition, we report a test of our misperception hypothesis in Experiment 4.

Conclusion

Overall, the evidence from Experiment 1 is mixed. Certain aspects of the results support the power PC theory and are clearly problematic for the contingency account and all variants of the RWM, whereas other aspects have the reverse implications. We provided two explanations for why deviations from causal power might have occurred. One was the ambiguity of the causal strength question asked, and the other had to do with potential failure to detect zero contingencies in a sequential trials design. The computational causal power account postulates that these deviations are artifacts of the experimental method used and unrelated to the underlying reasoning process. The competing approaches (i.e., RWM, Pearce, and linear models), in contrast, suggest that some or all of the deviations from causal power are rooted in robust aspects of the induction process and thus should prevail regardless of improvements in the experimental methods. The goal of Experiment 2 was to directly contrast these competing interpretations, by using a substantially improved experimental paradigm. We also included generative and preventive conditions in one single session, to address the criticisms raised by Lober and Shanks (2000).

Experiment 2

Experiment 2 included preventive, generative, and noncausal ratings in the same experimental setting. In particular, we measured causal judgments of pairs of candidates with (a) the same value of ΔP (positive, negative, or zero) as a function of the base rate of e, (b) the same causal power but varying values of ΔP, and (c) causal powers whose magnitudes had the opposite ordering as the absolute values of their ΔPs, therefore directly pitting the predictions of the two approaches against each other. Perales and Shanks (in press), using a design analogous to what we propose here, replicated the opposite influence of base rate on causal ratings when ΔP was kept constant. Thus, the RWM (and Lober & Shanks's, 2000, argument about different outcome saliencies in preventive and generative context) has already been refuted. Perales and Shanks also replicated the influence of base rate on noncontingent conditions and an influence of ΔP on conditions with equal causal power; they interpreted this as evidence against the power PC theory, but in favor of Pearce's (1987) model. However, these studies used the same ambiguous rating scale as we used in Experiment 1. The goal of Experiment 2 thus was to determine whether causal strength ratings from a substantially improved paradigm would still show these deviations from normativity that are problematic for the power PC theory, but follow from covariational models.

Measures to Rule Out Alternative Explanations of the Results

As we discussed, the question used to probe causal ratings in Experiment 1 was ambiguous. Rather than requesting a vague subjective estimate of causal strength on a rating scale, in Experiment 2 we asked participants to estimate the frequency of an outcome that was due to the candidate (e or $\neg e$, for generative and preventive candidates, respectively) under an intervention on entities that did not show that outcome. Estimating the frequency of a specific event is less likely to be confused with estimating reliability. In particular, we asked participants how many entities out of 100, all of which did not show an outcome, would now have the outcome in a counterfactual situation in which the candidate cause was introduced. Because the question concerns an intervention (see, e.g., Pearl, 2000), it requests a causal estimate; the counterfactual nature of the intervention justifies in an emphatic way the assumption that alternative causes in the context are constant, thereby allowing projections based on causal knowledge.

The context (before the intervention) we chose for estimating the extent of generative influence was one in which e never occurred, and the context for estimating the extent of preventive influence was one in which e always occurred. In these contexts, because alternative influences of e of like kind as the candidate (generative or preventive) are counterfactually removed, the influence of the candidate should manifest itself without contamination. That is, the estimated frequency of e in the (counterfactual) presence of the generative cause should reflect the strength of this cause alone; the same holds for the estimated frequency of $\neg e$ as a measure of the preventive strength of the candidate. There are no simpler or clearer contexts under which to manifest the strength of a candidate causal relation.

In addition, to encourage a uniform interpretation of the values on the dependent measure, we exposed every participant at the outset to a trio of candidates that should yield a full range of answers (Parducci, 1965). Specifically, within the first 3 screens, all participants were presented with a candidate that respectively had a positive, negative, or 0 ΔP.

One result derived in Cheng's (1997) theory is that causes in the context other than the candidate cause must occur independently of it (e.g., other causes are held constant) for the causal power of the candidate to be inferred. (This result is consistent with both intuition and empirical findings; see Cheng, 1997, for a review; also see Spellman, 1996.) It follows that to test this theory's predictions on estimated causal power, this precondition must be met. Unfortunately, the instructions in Experiment 1 may not have clearly conveyed this assumption, potentially contributing to the outcome-density bias.

To strongly encourage the assumption that alternative causes remain constant across the experimental and control contexts, our instructions explicitly informed the participants that in the studies they would see, patients were "randomly assigned" to two groups, one receiving a medicine and the other not receiving medicine. To promote reflection on this assumption, so that it can be sustained throughout the experiment, we asked participants a question regarding causal attribution in a noncontingent situation at the beginning of the experiment. They answered whether the outcome (headache, as a potential side effect of the medicine) in the group that received the medicine could be attributed to the medicine and justified their answer. There is no reason why our question could have blocked participants from attributing headaches in the experimental group to the medicine, except for the assumption that other causes remain constant across the two groups. Participants did not receive any feedback on their responses.

Our goal was to study people's natural capability to discover causal relations, with as little interference as possible from other mental processes. Demands on mental processes that could influence performance but are not directly relevant to causal inference would add ambiguity and statistical noise to our results. To reduce demands on comprehension, we used a visual format wherever we thought it would aid the clarity of presentation and reduce potential errors in comprehension. To reduce demands on memory, we presented all trials relevant to each candidate cause simultaneously, on the same screen. Finally, to keep boredom and fatigue to a minimum, we included as few conditions as possible that would still attain our goals, forgoing the exhaustive factorial combination of five levels of $P(e|c)$ and $P(e|\neg c)$ in Experiment 1.

Our approach to isolating the operation of the causal induction module was analogous to that of an everyday causal reasoner who attempts to isolate the influence of a candidate cause by eliminating the influence of alternative causes. The goal was to arrive at an explanation of behavior, rather than a mere circumstantial description of it.

Method

Participants

Fifty-one students from undergraduate psychology courses at the University of California, Los Angeles, participated to partially fulfill a course requirement.

Apparatus

Except for the paper response sheets, the experiment was presented on Macintosh computers using the Superlab Pro software (Cedrus, Phoenix, AZ).

Design

Manipulations of causal power and ΔP. There were 10 within-subject conditions (i.e., causal inference problems), each with a different pair of relative frequencies of *e* that allowed the estimation of $P(e|c)$ and $P(e|\neg c)$ respectively (see Table 3). For each condition, sample size was held constant at 72 trials, divided evenly between the two frequencies and presented simultaneously on the same computer screen. Each trial consisted of information on whether a candidate cause was present or absent and whether an effect in question was present or absent.

The 10 conditions are described below in sometimes overlapping pairs to clarify the underlying design. The conditions that involved generative candidates and those that involved preventive ones were mirror images of each other.

- a pair of conditions with the same ΔP of .50, but varying generative causal powers of .50 and .75 (Conditions D and B in Table 3), and its preventive analogue (Conditions G and E);
- a pair of conditions with the same generative causal power of .75, but varying ΔPs of .50 and .75 (Conditions B and C), and its preventive analogue (E and F);
- a pair of conditions with the same ΔP of 0 but differing base rates of *e* (1/3 and 2/3; Conditions H and I). The generative and preventive causal powers for these two conditions were 0; and
- two conditions with a causal power of 1, generative for one and preventive for the other (Conditions A and J). These conditions allowed the creation of (a) pairs of candidates for which ΔP and causal power make predictions of differences in opposite directions (Conditions A and C, and Conditions J and F) and (b) more candidates with the same ΔP value (namely, .50 or −.50) but varying causal powers.

In summary, some pairs of conditions varied the base rate of *e* for candidates with the same level of ΔP (positive, negative, or zero), yielding different causal powers for their members when ΔP was nonzero; other pairs manipulated ΔP for candidates with the same causal power. All participants received the same 10 conditions.

Table 3
Design and Results of Experiment 2

					Causal ratings	
Condition	Power	ΔP	$P(e\|c)$	$P(e\|\sim c)$	*M* (*SD*)	*Mdn*
A	1.00	0.50	36/36	18/36	85.7 (26.5)	100
B	0.75	0.50	30/36	12/36	67.8 (19.1)	75
C	0.75	0.75	27/36	0/36	74.4 (5.7)	75
D	0.50	0.50	18/36	0/36	48.5 (14.1)	50
E	0.75	−0.50	6/36	24/36	−59.7 (31.4)	−75
F	0.75	−0.75	9/36	36/36	−66.5 (18.8)	−75
G	0.50	−0.50	18/36	36/36	−44.5 (21.1)	−50
H	0.00	0.00	12/36	12/36	−0.5 (4.86)	0
I	0.00	0.00	24/36	24/36	0.7 (3.78)	0
J	1.00	−0.50	0/36	18/36	−85.9 (26.6)	−100

Note. The $P(e|c)$ and $P(e|\sim c)$ columns list how many of the sample of 36 patients in each respective group showed the effect in each condition. Preventive ratings are represented by negative numbers.

Measurement of participants' estimates of causal influence. The dependent variable, a participant's assessment of the strength with which the candidate cause influences the effect, was measured in two steps: (a) a qualitative assessment of whether a candidate had any influence on *e* and (b) an estimate of the extent of the influence if the candidate was judged to have an influence. This extent was indicated by a frequency of an outcome (*e* or ¬*e*) that was due to the candidate cause in a counterfactual situation involving an intervention with the candidate, as explained earlier.

Order of information and counterbalancing. As mentioned, we exposed all participants at the outset to a set of three candidate causes that were expected to yield a full range of answers on the dependent measure. These three candidates always had values of causal power in the following order: .50, 1, and 0, but we counterbalanced the ordering of the direction of causal power for the first 2 candidates between two groups (one group was presented with a generative cause before a preventive one; the other group was presented with a preventive cause before a generative one). We also counterbalanced the base rate of *e* for the third, noncausal candidate between the same two groups. The first 3 conditions were therefore D, J, and H for one group and G, A, and I for the other group. The remaining seven conditions for each group were presented in a different random order for each participant.

Materials and Procedure

Cover story. The 10 conditions were presented under the same cover story:

> You are an employee for a Company that distributes new medicines for preventing allergies. Your job is to review information regarding a possible side effect of the new allergy medicines that are under consideration for distribution. Although these medicines have been found to be clearly effective in preventing allergies, they may cause headaches, prevent headaches, or have no influence at all on headaches.
>
> You will see the results of experiments that were conducted to study the influence of these medicines on headaches. For each study, patients were randomly assigned to one of two groups: an experimental group that received the new medicine, and a control group that did not. Based on the data presented, judge whether each medicine has a side-effect on headaches, and if so, whether it causes or prevents them. Your success in the company is highly dependent on your accurate assessment of these side effects.

As indicated in the cover story, the candidate causes were various allergy medicines, and the effect was headache in patients in the study. A patient to whom a particular medicine had (or had not) been administered and who did (or did not) have a headache constituted a learning trial. Headache was presented as a possible side effect of the medicines to justify why potentially any subject sampled from the population of allergy patients could have been administered one of the medicines in question.

Encouragement to assume that alternative causes were held constant. To encourage reflection on the assumption that causes of headache other than a medicine in question occurred independently in the two groups, participants were asked to consider causal attribution in the following hypothetical situation before viewing the results of the various studies conducted:

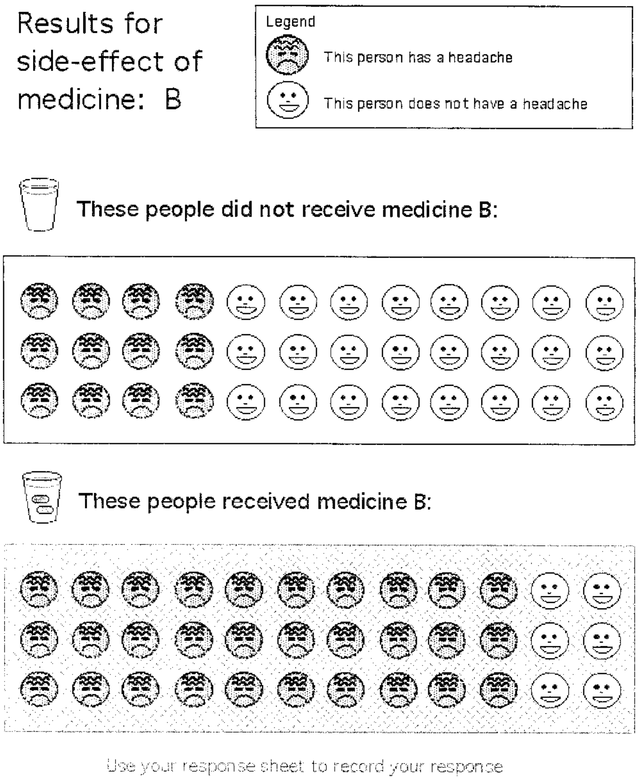
> We conduct a study of medicine X, and find that: 50% of the participants who received medicine X (those in the experimental group) have headaches. Likewise, 50% of the participants who did not receive medicine X (those in the control group) have headaches as well. Recall that participants were randomly assigned to the two groups.
>
> QUESTION: Can the headaches in the experimental group be attributed to medicine X?

Participants were asked to write down their answer and a justification for it. They received no feedback on their answers to this question.

Presentation of the 10 conditions. The participants were then instructed on the mechanics of the experiment, such as how to proceed from one screen to the next, following which they viewed the results of the 10 fictitious studies, each involving a particular medicine (denoted by a letter corresponding to its condition label). Recall that the first 3 studies were always presented in identical order in each of two groups whereas the remaining 7 were presented in random order.

All 72 trials that were potentially relevant to the evaluation of a medicine were presented simultaneously on the same computer screen (as illustrated in Figure 5). Each screen displayed information on how many patients in the respective groups—control and experimental—did and did not have headaches. The presence of the effect (headache) was represented by a blue frowning face and the absence of the effect by a white smiling face. Each of the 72 faces thus constituted a learning trial. The screen contained two panels of 36 (four rows of 12) faces each. The top panel had a white background and always represented patients from the control group who did not take the medicine under investigation; the bottom panel had a colored background and always represented patients from the experimental group who did take the medicine. In addition to the verbal labeling of the two panels (e.g., "These people did *not* receive medicine B" vs. "These people received medicine B"), a drawing of an empty glass accompanied the top panel, and a drawing of a glass containing two identically colored capsules accompanied the bottom panel. The background colors of the experimental group matched the color of the corresponding capsules. Ten highly differentiable colors represented the 10 different medicines under investigation.

Participants assessed the effect each allergy medicine had on headache. They first indicated whether they thought the medicine "has no effect on



Results for side-effect of medicine: B

Legend

This person has a headache

This person does not have a headache

These people did not receive medicine B:

These people received medicine B:

Use your response sheet to record your response

Figure 5. A black-and-white version of the stimulus materials in Experiment 2. The figure depicts the results of a study testing the influence of Medicine B on headaches.

headaches", "causes headaches", or "prevents or relieves headaches". If the participant responded that the medicine causes headaches, they were further asked to estimate how many out of 100 people who did not have headaches would have a headache if given the medicine. If the participant responded that the medicine prevented headaches, they were asked to estimate how many out of 100 people, all of whom had headaches, would not have a headache if given the medicine. If participants answered that the medicine "has no effect" on headaches, their answer was scored as 0. If they answered that the medicine "causes" or "prevents" headaches, the number they wrote in answer to the counterfactual question was scored verbatim (as a positive number); if they answered that the medicine "prevents" headaches, the subsequent number they wrote was scored as negative. All conditions used identical paper response sheets (participants had to fill in the label of the medicine they were evaluating, due to variation in the ordering of the medicines). In addition, participants were asked to briefly explain their answers.

Participants took approximately 25 min to complete the experiment and did so individually in separate rooms.

Predictions. Figure 6 displays predictions for all 10 conditions of Experiment 2 for each of the following models: the power PC theory, the unrestricted RWM with $\beta_{US} < \beta_{\overline{US}}$ and $\beta_{US} > \beta_{\overline{US}}$ and Pearce's (1987) model. As before, predictions for the power PC theory were simply derived by inserting the values of ΔP and $P(e|\neg c)$ into the relevant equations. Whether Equation 3 or 4 applied was entirely dependent on the sign of ΔP.[7] To derive predictions for the RWM, we used the same parameter settings as in Experiment 1. Asymptotic predictions for the Pearce model were based on a formula adapted from Perales and Shanks (in press):

$$J = x_2 \frac{a\lambda(c + d) - c\lambda(ax_1 + bx_1)}{a(c + d) + b(c + d) - cx_1(ax_1 + bx_1) - dx_1(ax_1 + bx_1)}, \tag{8}$$

where a through d represented the cells of a contingency table, x_1 reflected the similarity between the context in isolation (X) and the cause-context compound (CX) and x_2 reflected the similarity between the cause in isolation (C) and CX. Both x_1 and x_2 were determined by the saliencies of X and C, S_x and S_c respectively, with

$$x_1 = \frac{S_x}{S_x + S_c}, \tag{9}$$

and

$$x_2 = \frac{S_c}{S_x + S_c}. \tag{10}$$

We based our predictions on the assumption that C is four times more salient than X, yielding values of .2 for x_1 and .8 for x_2.[8]

In Figure 6 generative causal strengths are plotted as positive values and preventive causal strengths as negative values.[9] A visual comparison of Panels A through D reveals that none of the purely covariational models predicts the same pattern as does the power PC theory. The RWM cannot predict an interactive influence of the base rate of e and the sign of ΔP on causal ratings for candidates with identical levels of nonzero ΔP, irrespective of its parameter settings. Pearce's (1987) model can predict this interaction, but (unlike either the power PC theory or the RWM) it also predicts that (a) noncontingent candidates should elicit nonzero causal ratings that are sensitive to variations in $P(e|\neg c)$ and (b) all conditions with the exception of Condition J, including those with negative ΔPs, should have positive associative strengths. Although Pearce's model predicts a slight influence of ΔP on conditions with equal power, compared with RWM this influence is negligible. We come back to the issue of positive associative strengths for candidates with negative ΔPs in the General Discussion section.

Results and Discussion

Causal Judgments: Paired Comparisons

The right columns of Table 3 display the mean and median causal ratings for the 10 conditions. We report the medians in addition to the means because the frequency distributions for most conditions were clearly skewed. Figure 7, a histogram for Condition E, illustrates a distribution that is moderately skewed among the range observed in this experiment. One reason for the increased skewness in this experiment relative to Experiment 1 was that new rating scale included directionality, resulting in an increase of the range from 100 to 200 points; a slippage on the direction of influence by a single participant, for example, could therefore shift the mean ratings substantially. Comparing the obtained means and medians with the predictions derived from the power PC theory (Column 2 in Table 3) reveals that our clarification of the materials yielded observations that clearly favor the causal power approach over the purely covariational approaches. We used paired sign tests for all analyses.[10] In contrast to Experiment 1, in which we

[7] For the two noncontingent conditions, either equation can be applied; causal power always is zero.

[8] It appears that Perales and Shanks (in press) have added an extra degree of freedom into the Pearce model by tacitly assuming that the similarity relation between CX and X is asymmetrical (i.e. CX is more similar to X than X is to CX). It is well known that similarity relations often are asymmetrical (see, e.g. Tversky, 1977). Pearce's conception of similarity based on saliency does not allow asymmetries, however (see Equation 9; Pearce, personal communication, February 2003). How Perales and Shanks arrived at the similarity parameters that best fit their data is therefore not clear, nor are the implications of their results for Pearce's theory.

[9] In Figure 6, because preventive causal strength increases downward (because of their representation as negative numbers on the y-axis), decreases in preventive strength for conditions with identical ΔP as $P(e|\neg c)$ increases are represented by positively sloping lines. The power PC theory's prediction of an interactive influence of $P(e|\neg c)$ and the sign of ΔP on causal ratings for candidates with identical ΔPs can most readily be seen by considering the absolute magnitudes of deviations from 0, the horizontal line across the middle. This interaction corresponds to positive slopes in lines in the figure connecting preventive candidates as well as those connecting generative ones.

[10] The skewness of the distributions of causal estimates renders the interpretation of the means unclear; the interpretation of statistical tests of a difference between means is therefore likewise unclear. Because the dependent measure in this experiment (estimated number of patients) is on a ratio scale, transformations of the scale to remove the skewness would be undesirable. We report the p levels of the sign tests because they are more meaningful. Other analyses show the same general pattern of statistical reliability as the sign tests at the α level of .05. Planned comparisons in a Candidate Cause (10 medicines) × Order of Initial Candidates (Medicine D vs Medicine G first) analysis of variance, with candidate cause as a within-subject factor and order as a between-subjects factor, show the exact same pattern of reliability. Paired t-tests show the same pattern except for the comparison between Conditions B and C, with $t(50) = 2.56$, $p < .02$. Because of the extreme skewness of the distributions, however, this result does not reflect what most of the participants were doing: although the mean of Condition C is higher than that of B, more participants gave B a higher rating than C. It is possible that some subjects responded according to ΔP; an analysis of this hypothesis is presented later.

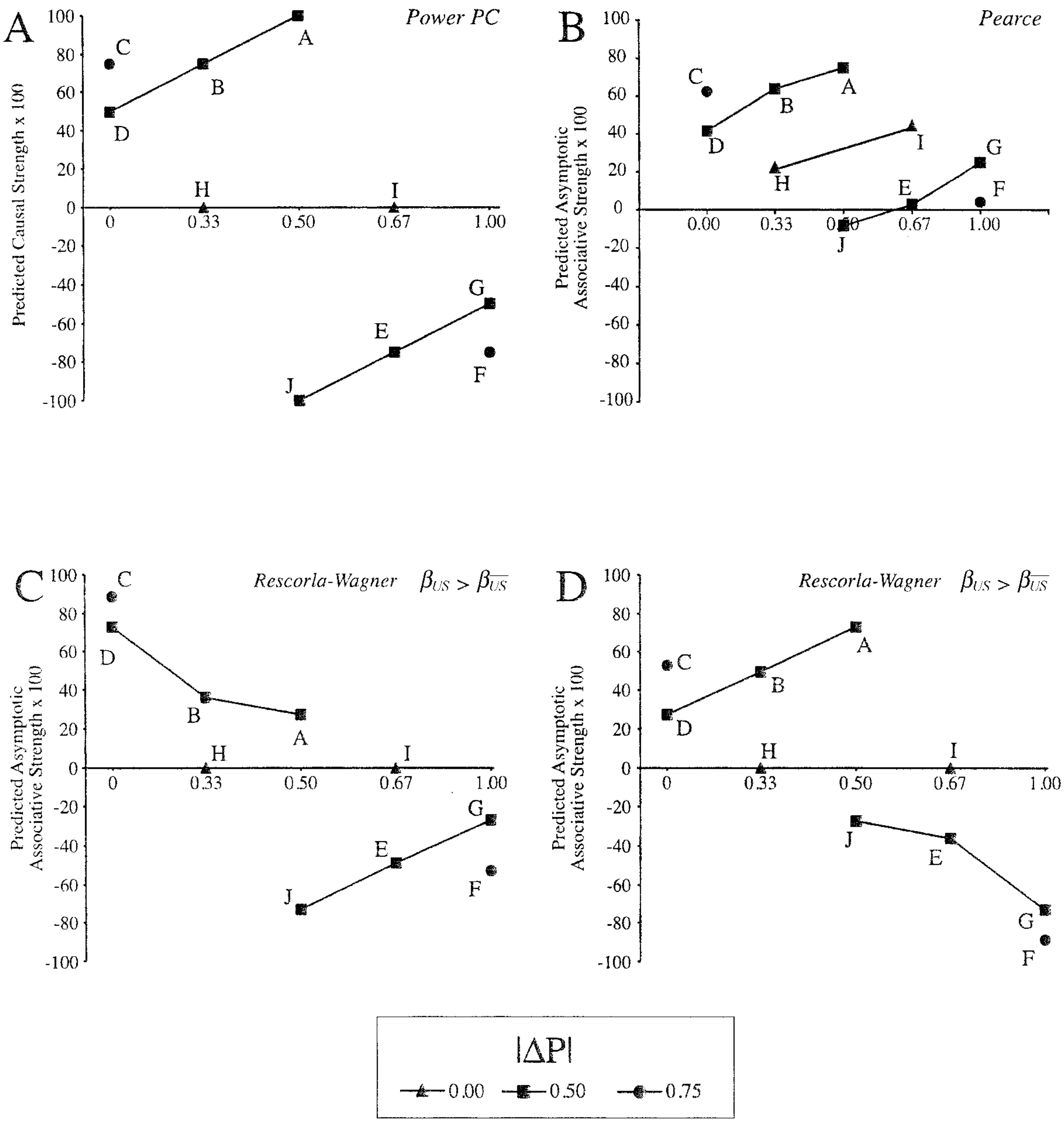


Figure 6. Predicted causal estimates for Experiment 2 as a function of $P(e|\neg c)$ made by the power PC theory (A), Pearce's (1987) model (B), and the Rescorla-Wagner (1972) model with $\beta_{US} > \beta_{\overline{US}}$ (C) and $\beta_{US} < \beta_{\overline{US}}$ (D). Lines connect conditions with identical levels of ΔP; the abscissa represents $P(e|\neg c)$. Letters A through J denote the corresponding conditions for each point (see Table 3). Note that Conditions B and C have identical causal powers, as do Conditions E and F. Panel A also depicts the medians of the observed causal ratings in Experiment 2.

investigated trends in causal ratings for several conditions sharing identical levels of ΔP or causal power, we designed Experiment 2 to allow paired comparisons between maximally informative conditions.

Same ΔP, different causal powers for contingent candidates only. Inspection of Table 3 reveals that the differential predictions of the power PC theory are well supported. Causal judgments (reflected by the mean and median frequency estimates) in conditions with identical positive, negative, or zero ΔP are a complex function of the base rate of *e* (the sixth and seventh columns in the table). In particular, with an increasing base rate of *e,* the absolute values of the average causal judgments of candidates with the same level of ΔP (a) increased when ΔP was positive (cf. Condition D with Condition B, and Condition B with Condition A; see Table 3), (b) decreased when ΔP was negative (cf. Condition E with Condition G, and Condition J with Condition E), and (c) remained unchanged at or near 0 when ΔP was 0 (cf. Condition H with Condition I). This pattern of results contra-

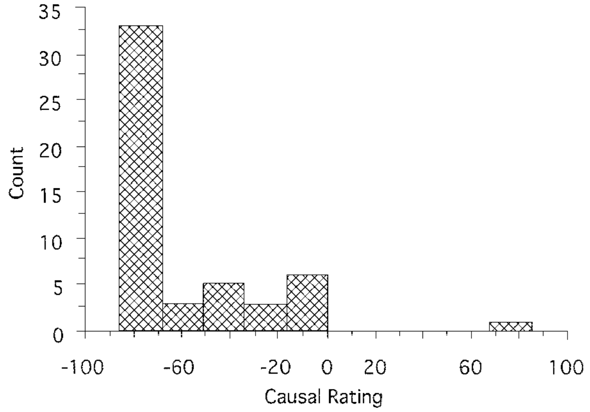


Figure 7. Histogram of Condition E in Experiment 2: an example of a moderately skewed distribution of causal judgments.

dicts the traditional ΔP model and all consistent parameter settings of all purely covariational models. For simplicity, we first restrict our discussion of the implications of our results to those for the RWM.

Paired sign tests of these comparisons indicated that this qualitative pattern was highly reliable. Conditions A (causal power = 1.00), B (causal power = .75), and D (causal power = .50) shared the identical positive ΔP = .50. Paired sign tests between Conditions A and B (number who rated A higher than B = 38; number who rated B higher than A = 11; number of ties = 2; z = 3.70, $p <$ .001) and Conditions B and D (number who rated B higher = 35; number who rated D higher = 4; number of ties = 12; z = 4.80, $p <$ .001) revealed that reliably more participants rated in accordance with causal power. Analogously, Conditions J (preventive power = 1.00), E (preventive power = .75), and G (preventive power = .50) shared the identical negative ΔP = −.50. For these conditions, paired sign tests between Conditions J and E (number who rated J as having higher preventive power than E = 41; number who gave the reverse ordering of ratings = 9; number of ties = 1; z = 4.40, $p <$ .001) and E and G (number who rated E with a higher preventive power than G = 37; number who gave the reverse ordering of ratings = 10; number of ties = 4; z = 3.80, $p <$ .001) also showed that reliably more participants judged in accordance with causal power.

ΔP was 0 for Conditions H and I, both of which nearly always elicited a judgment that the candidates had no effect (this was true for 47 out of the 50 participants who gave a rating for both; 1 participant omitted to judge Condition I). The mean judgments in these conditions were independent of the base rate of e, as indicated by a sign test using a binomial distribution (number who rated H higher = 0; number who rated I higher = 2; number of ties = 48; p = .50).

Same causal power but different Δ*Ps.* To further test the causal power approach against purely covariational accounts, we compared pairs of conditions with identical causal powers but different ΔPs. As Table 3 shows, the medians of causal judgments for candidate causes in such pairs were identical, supporting the power PC theory but contradicting the traditional ΔP and variants of the RWM. Corroborating this conclusion, the mode and median of the within-subject differences between candidates in each such pair was 0. Conditions B and C, and E and F, all shared generative and preventive causal power of .75, respectively. The difference in ΔP (.50 vs. .75) did not influence causal ratings in Conditions B and C (number who gave B a higher rating = 20; number who gave C a higher rating = 19; number of ties = 12; z = 0.00, $p >$ .50). Analogously, different negative ΔPs (−.50 vs. −.75) did not affect causal ratings in Conditions E and F (number who rated E with a higher preventive strength = 13; number who rated F with a higher preventive strength = 16; number of ties = 22; z = 0.71, $p >$ .50).

Directly pitting causal power against Δ*P.* Recall that for two pairs of conditions, A and C, and their preventive analog, J and F, causal power and ΔP predicted differences in opposite directions. Results for these pairs further supported the power PC theory: The condition in each pair with the higher causal power (A and J) reliably received higher causal estimates more often, despite its lower (absolute) ΔP (for the AC pair: number who rated A higher than C = 39; number who gave the reverse ordering of ratings = 11; number of ties = 1; z = 3.82, $p <$ .001; for the JF pair: number who rated J with a higher preventive strength than F = 40; number who gave the reverse ordering of ratings = 11; number of ties = 0; z = 3.92, $p <$ .001).

Can Alternative Covariational Accounts Explain the Data From Experiment 2?

The results of Experiment 2 also clearly refute other purely covariational accounts (see Figure 8). The weighted ΔP model (e.g., Anderson & Sheu, 1995) and linear combination models (Schustack & Sternberg, 1981), with their typical ordering of parameter weights (consistent with that for our data), all erroneously predict that (a) candidates with identical causal powers should be affected by variations in ΔP and (b) the ordering of causal estimates for members of the two pairs of conditions, A and C and J and F, should be opposite to the ordering of their causal powers (recall that causal power and ΔP were directly pitted against each other for these pairs). These models and Pearce's (1987) model further predict that noncontingent candidates should elicit different judgments depending on the base rate of e. These three predictions, all disconfirmed in Experiment 2, are also made by Van Hamme and Wasserman's (1994) reformulation of the RWM.

Although a range of purely covariational and associationist models could account for the particular pattern of results from Experiment 1, including aspects of the results that were at variance with the predictions of the power PC theory, these approaches are all refuted by the data from Experiment 2. Our hypothesis—that the deviations from normativity observed in Experiment 1 were the result of ambiguities and performance limitations and did not reflect fundamental aspects of the inductive process—was confirmed. When these problems were curtailed, causal judgments were normative throughout, as predicted by the power PC theory. No purely covariational model or associative learning theory, regardless of its parameter settings, can account for the results obtained in Experiment 2.[11]

[11] Our interpretation that the variation in causal judgments observed in Experiment 1 for candidates with the same nonzero causal power was due to the ambiguity of the questions is corroborated by work in progress from Collins and Shanks (personal communication, April 2002), showing that

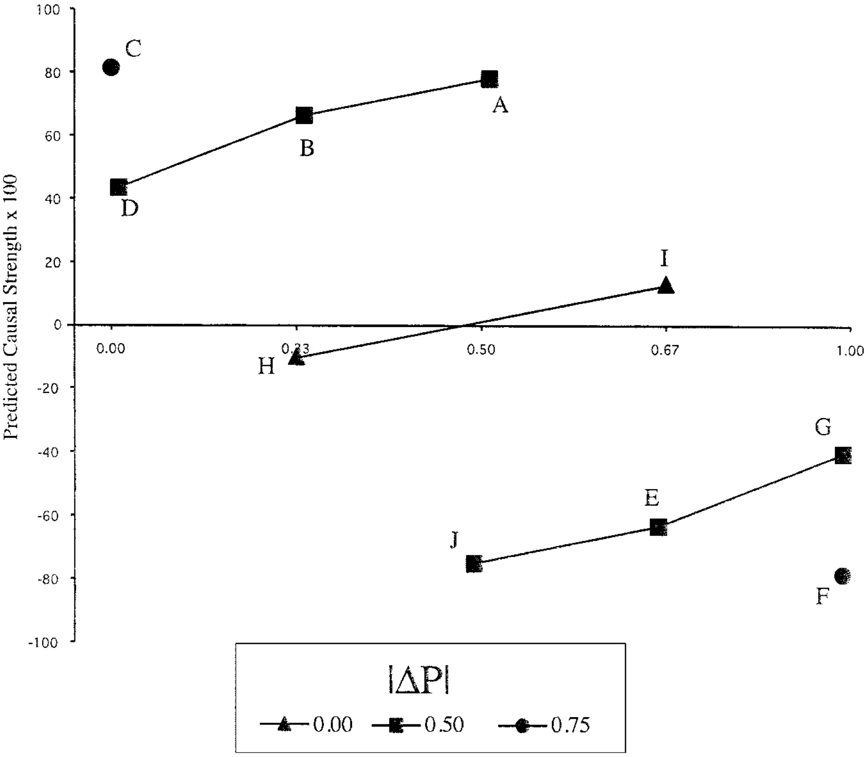


Figure 8. Best fit of the linear combination model to the causal estimates observed in Experiment 2 as a function of $P(e|\neg c)$. The weights were as follows: $w_a = 4.243$, $w_c = -2.316$, and $i = -33.189$.

Judgment Strategies

In addition to contrasting individual conditions against each other, it is also informative to investigate what strategies underlie participants' general rating behavior. Participants could, for example, base their ratings on causal power or ΔP or they could use simpler strategies such as rating according to $P(e|c)$ (cf. Klayman & Ha, 1987). Given the robust finding of a split between participants who generally rated according to either ΔP or causal power we found in Experiment 1, such an analysis would be particularly illuminating. We analyzed judgment strategies consistently used by each participant throughout the 10 conditions and classified each participant's ratings as driven by a particular strategy if all ratings were within a margin of 10 points from the respective ratings predicted by that strategy for each of the 10 conditions. The strategies included in our analysis were as follows: causal power, ΔP, Cell A (equivalent to $P(e|c)$ for our materials), and base rate of *e*. This analysis revealed that 24% of the participants consistently based their judgments on causal power in all 10 conditions. Of the remaining participants, 74% used causal power for more conditions than any other strategy, resulting in an overall preference for causal power in 80% of the participants.[12] In contrast, not a single participant consistently used any of the other strategies. This qualitative analysis of judgment strategies corroborated our conclusion drawn from pairwise comparisons of maximally informative conditions.

Experiment 3

All aspects of the predictions derived for the power PC theory were well supported by the data in Experiment 2. The opposite influence of base rate on causal ratings for positive and negative contingencies prevailed despite the same scenario covering all contingencies, refuting Lober and Shanks's (2000) argument that different cover stories in Experiment 1 resulted in a reversal of outcome saliencies. In contrast, the deviations from causal power did not prevail when methods were used to reduce the ambiguities in the materials. The observed pattern of results cannot be explained by any of the alternative covariational or associationist models.

One might argue, however, that the method and data from Experiment 2 fall outside the scope of associative learning theory, because we presented the trials simultaneously on the computer

variations of question format (counterfactual vs. standard scale) mediate the degree of accordance with causal power versus ΔP. Studies such as Collins and Shanks's are clearly informative. Our goal in Experiment 2 differed from theirs, however. Our goal was to study the process of causal inference, removing ambiguities in the interpretation of the findings due to extraneous influences as much as possible, rather than to test the effect on performance of historically used experimental procedures per se. We therefore did not independently manipulate each experimental procedure we introduced.

[12] The remaining 20% is distributed as follows: "tie" between ΔP and causal power, 12%; preference of ΔP, 6%; preference of Cell-A, 2%.

screen, preventing any operation of an error-correcting algorithm, which underlies many associative learning models. However, because no time limit was imposed on how long the materials remained on the screen, it was possible for participants to apply an error-correcting algorithm to each trial in turn. Therefore, there was no reason why associative learning could not have taken place in such paradigms, and indeed several researchers in the associative learning community have applied the RWM to predict causal judgments from stimuli presented in list form (see, e.g., Van Hamme, Kao, & Wasserman, 1993). If associative learning is the primary causal learning mechanism that is applied whenever it is possible, one would expect participants to scan each item in turn and engage in error correction. Therefore, the results observed in Experiment 2 disconfirm this role for associationist models.

It is possible, however, that multiple causal learning mechanisms were available, and participants in Experiment 2 chose a nonassociationist version, perhaps because it was more efficient for simultaneously presented trials. To test the generalizability of our results, we replicated Experiment 2 but used a sequential trials learning procedure, retaining as much as possible other aspects that clarified the task and avoided statistical noise.

Method

Participants

Thirty-one students enrolled in undergraduate psychology courses at the University of Sheffield, England, participated to partially fulfill a course requirement.

Apparatus and Design

The design was identical to Experiment 2, but the stimuli were presented on a trial-by-trial basis on the computer screen. The software used to create the experiment was PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). To avoid fatigue and shorten the overall duration of the experiment, the stimulus set was reduced to 24 trials per condition. Participants used the keyboard to enter causal ratings.

Materials and Procedure

Participants read the same cover story as in Experiment 2, but a clause was added stating that the studies were conducted in different countries. This was done because some participants from Experiment 2 mentioned in the debriefing that they were puzzled about the changes in base rate across conditions. After participants read the cover story, they proceeded to a screen explaining the procedure of the experiment (i.e., that there were 10 conditions and that they would study individual lab reports for each patient, etc.). They then viewed eight demonstration trials, two for each contingency cell a, b, c, and d. After they had provided a causal rating for these trials, they saw a screen that encouraged reflection on the assumption that alternative causes are constant (see Experiment 2). Once they had answered (Y)es or (N)o, they proceeded to the 10 experimental conditions.

In each learning trial, the computer displayed a line on the top of the screen stating which medicine (A through J) was currently under investigation. Centered on the left was displayed either an empty glass (cause absent) or a glass containing two colored capsules (cause present), analogous to Experiment 2. In addition, a verbal description (e.g., “This patient did/did not take medicine X”) was displayed above the picture. The outcome was displayed 1 s later. Centered on the right was displayed either a smiling or a frowning face, accompanied by a verbal description above the picture (“. . . and has a/no headache”). Five hundred ms after the outcome presentation, a prompt asked participants to press SPACE to proceed. After participants had observed all 24 trials of a condition, they were asked to rate the candidate’s causal strength. Analogous to Experiment 2, there first was a qualitative rating, where participants indicated by pressing a key whether the candidate in question (C)auses or (P)revents headaches or has (N)o impact on them. If participants chose causes or prevents, they then had to provide a quantitative estimate of causal strength. The same questions as in Experiment 2 were used.

The order of the first 3 conditions was counterbalanced between participants as in Experiment 2, and the remaining 7 conditions were presented in a different random order for each participant. The 24 learning trials within each condition were also presented in random orders. For the participants the medicines were always labeled A through J, and this order was used sequentially, irrespective of the actual condition label (as referred to in Table 3).

Results

Mean and median causal ratings for the 10 conditions are listed in Table 4. A comparison between Tables 3 and 4 reveals that Experiment 3 overall replicated the general pattern of results in Experiment 2 but that the data were noisier. Analogous to Experiment 2, we conducted nonparametric pairwise comparisons and adopted a significance level of .05.[13]

Same ΔP but Different Causal Powers

Participants reliably rated Condition A higher than Condition B (number who rated A higher = 22; number who rated B higher = 8; 1 tie; $z = 2.37$) and Condition B higher than D (number who rated B higher = 21; number who rated D higher = 8; 2 ties; $z = 2.23$). Likewise, participants rated Condition J higher than E (number who rated J higher = 22; number who rated E higher = 6; 3 ties; $z = 2.84$) and E higher than G (number who rated E higher = 20; number who rated G higher = 6; 5 ties; $z = 2.55$).

Noncontingent Conditions

Contrary to Experiment 2, the two noncontingent conditions, H and I, seemed to elicit different causal ratings ($z = 2.41$). A closer examination of the data distribution reveals that in fact the majority of participants ($n = 20$) rated both conditions as noncausal. However, a sizeable minority ($n = 6$) provided medium to strong preventive ratings for H (*Mdn* = 65.0) while providing either noncausal or medium to strong generative ratings for I, and another minority ($n = 4$) provided medium to strong generative ratings for I (*Mdn* = 65.0) while providing noncausal ratings for H resulting in a significant difference on the paired sign test. (One participant provided a weak generative rating for H, while rating I to be noncausal.)

Same Causal Power but Different ΔPs

Condition C did not receive significantly higher ratings than B despite C’s higher value of ΔP (number who rated B higher = 10;

[13] Planned comparisons in a two-way (Candidate Cause × Order) ANOVA and paired *t* tests show the exact same pattern of reliability as the reported sign tests, with the exception of the comparison between Conditions J and F, which was significant on a *t* test, $t(30) = 2.31$. The power PC theory predicts a difference between these conditions in the observed direction.

Table 4
Causal Ratings in Experiment 3

			Causal ratings	
Condition	Power	ΔP	*M* (*SD*)	*Mdn*
A	1.00	0.50	67.6 (38.9)	80
B	0.75	0.50	48.1 (33.5)	60
C	0.75	0.75	59.2 (26.0)	70
D	0.50	0.50	20.8 (30.4)	20
E	0.75	−0.50	−49.4 (32.4)	−60
F	0.75	−0.75	−53.4 (33.8)	−70
G	0.50	−0.50	−22.1 (32.2)	0
H	0.00	0.00	−12.3 (27.0)	0
I	0.00	0.00	10.8 (28.3)	0
J	1.00	−0.50	−73.2 (32.0)	−90

number who rated C higher = 17; 4 ties; z = 1.16, p = .25). Likewise, F did not receive significantly higher ratings than E (number who rated F higher = 15; number who rated E higher = 11; 5 ties; z = 0.59, p = .56).

Directly Pitting Causal Power Against ΔP

The two pairs of conditions, A and C (number who rated A higher = 20; number who rated C higher = 10; 1 tie; z = 1.64) and their preventive analogs, J and F (number who rated J higher = 20; number who rated F higher = 10; 1 tie; z = 1.64) elicited causal ratings in the direction predicted by causal power, rather than ΔP. Unlike in Experiment 2, however, these differences were not significant, probably because of the noisier data.

Discussion

Overall, Experiment 3 replicated the main findings of Experiment 2. As would be expected from a memory-taxing, sequential trials design, the data were considerably noisier compared with Experiment 2, but the crucial findings were still obtained: Participants based their ratings on causal power, both in situations where ΔP was constant (and causal power varied) and when ΔP varied and power was constant. Contrary to Experiment 2, however, but similar to Experiment 1, there was an outcome density bias in the noncontingent conditions.

Given that the majority of participants correctly estimated both noncontingent conditions to be noncausal, it is prudent to examine why a minority of participants might have deviated from normativity. As we hypothesized for Experiment 1, it is possible that these participants failed to perceive Conditions H and I to be noncontingent but in fact perceived a very small positive or negative contingency (as is more likely in a sequential trials design). Consider participants who erroneously perceived a zero contingency to be positive in Condition H or I. According to the power PC theory, judgments of generative power increase as the base rate of e increases for candidates with the same positive ΔP. Therefore, from the positive misperceptions, one would expect higher positive ratings for Condition I compared with H (I has a higher base rate than H). Conversely, if a zero contingency is misperceived to be negative, judgments of preventive power should decrease as base rate increases for candidates with the same misperceived negative ΔP; consequently, from the negative misperceptions of ΔP, one would expect weaker preventive ratings for I than for H (i.e., more headaches occurring for I than for H). Thus, misperceptions in either direction would contribute to an outcome-density bias.

Judgment according to causal power can explain not just the outcome-density bias but also the particular pattern of deviations from noncausality we observed. Recall that whereas most of the noncausal ratings for H were preventive, most of those for I were generative. Elaborating our explanation, if one assumes a threshold for reporting that a causal power is nonzero, so that candidates with causal powers that are too small are treated as noncausal, more participants would give preventive causal ratings in Condition H than I (recall the weaker preventive ratings expected for I than for H), whereas more participants would give generative causal ratings in Condition I than in H (recall the stronger generative ratings expected for I than for H).

Our interpretation of the outcome-density bias in terms of causal power is of course post hoc and hinges on the assumption that some participants did indeed misperceive ΔP. A simpler alternative explanation is that participants accurately perceived ΔP to be 0, but some chose to base their causal judgments solely on the outcome in the experimental patients, in other words, $P(e|c)$, treating the control group as irrelevant. Yet another explanation is that the frequency information was accurately perceived, but the causal inference process intrinsically yielded the outcome-density bias (e.g., Pearce, 1987; Schustack & Sternberg, 1981), as we discussed earlier. In fact, the results from Experiment 3 are perfectly in line with ordinal predictions derived from Pearce's model (see Figure 6). Experiment 4 aimed to shed light on the outcome-density bias for situations with memory demands that might lead to misperceptions of ΔP.

Experiment 4

Experiment 4 is a replication of Experiment 3, but in addition to prompting for causal estimates, we wanted to selectively check participants' perceptions of the contingencies in the noncontingent conditions. To this end, we introduced a twist to the cover story: Participants were told that it is the company's policy to conduct random spot-checks on the accuracy of their employee's assessments. Whether a spot-check occurs would be completely independent of their accuracy and reliability. In a spot-check, participants had to indicate whether headaches occurred more often, equally often, or less often in the experimental group they have just studied. Spot-checks were presented after rather than before the causal rating so as not to influence the rating. If our hypothesis that the observed outcome-density bias for noncontingent conditions in Experiments 1 and 3 was due to erroneously perceived contingencies, then this should be reflected in the spot-check answers.

Method

Participants

Sixteen students enrolled in undergraduate psychology courses at the University of Sheffield, England, participated to partially fulfill a course requirement or to receive £1 (approximately $1.59).

Apparatus, Design, Materials, and Procedure

All aspects of the experimental design and procedure were identical to Experiment 3, apart from the introduction of the spot-check component.

After the initial instructions and demonstration trials as in Experiment 3, the spot-check story was introduced as follows:

> The company you work for has the policy to check how careful its employees review the laboratory records. These checks are random spot-checks and whether or not you are being checked is completely independent of your accuracy or reliability. When you have been selected for a spot check, this means that you will be asked a question about the most recent set of laboratory records you have studied:
> For medicine X, did headaches occur
> (M)ore, (L)ess, or (E)qually often in the experimental group (those who took medicine) than in the control group?
> Your answer:
> Patients who took medicine X had headaches
> —MORE often than (M)
> —LESS often than (L)
> —EQUALLY often as (E)
> patients who did not take the medicine.

After participants provided their answer, the experiment proceeded as in Experiment 3, but with random spot-checks occurring after Conditions I and H and after Conditions B and E. Spot-checks after Conditions B and E were introduced to avoid selectively marking the noncontingent conditions with spot-checks.

Results and Discussion

The pattern of causal ratings replicated the findings from Experiment 3. For the sake of brevity, we only report the findings relevant to the noncontingent conditions. As in Experiment 3, some participants correctly rated both Conditions H and I to be noncausal, but others did not. Of the 9 participants whose spot-checks indicated at least one perceived nonzero ΔP, 2 rated Condition H higher than I, and 7 rated I higher than H ($z = 1.33$, $p = .18$); the other 7 participants indicated that an equal number of patients had headaches in the two groups for each condition and rated both conditions as noncausal with *more, less,* and *equally often,* respectively paired with positive, negative, and zero causal ratings. In accordance with our causal power explanation, the spot-check answers mirrored the causal ratings.[14] Not a single participant exhibited the pattern of results that follows from models that predict an outcome-density bias: perceived noncontingent input coupled with nonzero causal judgments. These results show that when participants judged the noncontingent candidates as causal, their judgments were not due to a hypothesized intrinsic aspect of the process of causal inference, such as the differential weighing of accurately perceived frequency information, nor due to responding on the basis of $P(e|c)$ alone, but entirely due to a misperception of the frequencies.

As would be expected, about the same proportion of participants misperceived ΔP to be nonzero in the two conditions (5 for H and 4 for I). Replicating the pattern of noncausal ratings in Experiment 3, for Condition H most of these participants (4 out of 5) provided medium to strong preventive ratings (−30 to −70), whereas for Condition I, most of them (3 out of 4) provided medium to strong positive ratings (20–85). This pattern further corroborates our causal power explanation and contradicts those by alternative accounts. The observed outcome-density bias for noncontingent conditions thus can indeed be attributed to a subset of participants who erroneously perceived a nonzero contingency. In other words, there was no actual outcome-density bias.

GENERAL DISCUSSION

Summary

We experimentally contrasted two approaches to causal inference: the causal power approach (Cheng, 1997) and the purely covariational approach, the latter with variants that included the classic contingency model (Jenkins & Ward, 1965), the weighted ΔP model, a benchmark associative learning theory (Rescorla & Wagner, 1972), Van Hamme and Wasserman's (1994) modification of the RW, Pearce's (1987) model, and linear combination models (Schustack & Sternberg, 1981). These approaches differ in their predictions (see Figures 1 and 6), notably for situations in which (a) ΔP varies but causal power remains constant; (b) a nonzero ΔP is constant, but the base rate of *e* varies, yielding different causal powers; and (c) $\Delta P = 0$, yielding a causal power of 0 regardless of the base rate of *e* whenever causal power has a defined value.

All four experiments reported here demonstrated a substantial influence of the base rate of *e* across conditions in which a nonzero ΔP remained constant. This influence was diametrically opposite for candidate causes with positive versus negative ΔPs. Experiment 2, which removed several plausible methodological problems in Experiment 1 and other studies (e.g., Lober & Shanks, 2000; Perales & Shanks, in press), yielded a surprisingly normative picture of human causal induction: There was no longer any influence of the base rate *e* when ΔP was zero or any influence of ΔP when causal power was constant. The complex pattern of results just described was as predicted by the parameter-free power PC theory but contradicts all purely covariation-based accounts of causal inference. The results suggest that the mental leap from covariation to causation requires that the reasoner hold the conviction that (unobservable) causal powers exist in the environment and that the goal of causal induction is to infer them.

Interaction Between the Sign of ΔP and the Base Rate of *e* Is Independent of Outcome Salience

Lober and Shanks (2000), in a critique of a preliminary report of the data from Experiment 1 (Buehner & Cheng, 1997), argued that a reversal in the RWM's parameter settings between scenarios can justifiably account for the opposite influences of the base rate of *e* for generative and preventive candidates with the same nonzero ΔP observed in Experiment 1. They argued that such a reversal reflects a difference in outcome salience across the preventive and generative scenarios we used in that experiment.

Experiments 2 through 4 tested Lober and Shanks's (2000) argument directly by including both generative and preventive candidates under one scenario, ensuring that the outcome was equally salient for positive, negative, and zero contingencies. The previously observed interaction between the influence of the base rate of *e* and causal direction was still obtained, empirically contradicting Lober and Shanks's argument (see also Perales & Shanks, in press, for a similar finding).

[14] There was one exception: One participant indicated that she perceived a positive contingency (Condition I), at the same time that she rated that relation as noncausal.

Explaining Deviations From the Predictions Based on Causal Power Observed in Experiment 1

Not all of the results from Experiment 1 supported the power PC theory. We proposed that these deviations can be explained by ambiguities in the causal question and memory limitations. In support of our hypotheses, when these ambiguities were curtailed, causal judgments were no longer a function of ΔP when causal power was constant. The approximately even split of participants between those who generally based their ratings on either ΔP or causal power observed in Experiment 1 disappeared when the materials did not allow more than one valid interpretation of the rating instructions. Causal power became the strategy preferred by the overwhelming majority of participants; in fact, it was the only strategy used consistently across conditions.

This normative pattern of results was replicated in Experiment 3, which adopted a traditional, sequential trials procedure to test the generalizability of our findings. The sequential trials procedure, however, used in Experiment 3 did produce an outcome-density effect on causal ratings in a minority of participants. Informed by a post hoc analysis of the subjective probability estimates from Experiment 1, we hypothesized that the underlying reason for this bias was that participants erroneously perceived a contingency when in fact there was none. If one makes this assumption, the power PC theory would predict an outcome-density bias exactly in the form observed in Experiment 3. Experiment 4 confirmed our hypothesis: Participants who provided causal ratings for noncausal candidates did so because they erroneously thought there was a contingency between cause and effect.

The Pearce (1987) Model Applied to Human Causal Learning

The assumption that the candidate Cause *C* is more salient than the Context *X* produced the best qualitative fit of the Pearce (1987) model to the data; in fact, the data from Experiment 3 match all ordinal predictions made by the Pearce model. However, even though the predictions are ordinally correct, they are at sharp variance with another qualitative aspect of the data: the sign of the accrued associative strengths. The four conditions involving negative contingencies (F, G, E, and J) all produced moderate to strong observed preventive (i.e., negative) causal ratings, yet the Pearce model predicts weak to moderate positive strengths in all of these conditions except J. The claim that the Pearce model can successfully predict the interactive influence of the sign of ΔP and the base rate of *e* on human causal ratings (cf. Perales & Shanks, in press) thus cannot be maintained.

The predicted pattern of little or no associative strengths, and in particular no negative associative strengths for negative contingencies, stems from the combination of the saliency assumptions we made, which were necessary to obtain the best ordinal fit to the data, and the surprise-based nature of associative learning algorithms. Error correction (i.e., changing associative strengths) only takes place when an unexpected outcome occurs; if there is no surprise, no error has to be corrected, and hence nothing new is learned. Under the assumption that *C* is more salient than *X*, any generalization between *CX* and *X*, as indexed by x_1 in Equation 8, is very small. In conditions with negative contingencies, the outcome occurs more often in the presence of the *X* alone than with *CX*. Consequently, *X* accrues excitatory strength, reflecting the fact that the outcome apparently occurs on its own, in the absence of *C*. But, because very little generalization between *X* and *CX* takes place, the outcome would not be expected on *CX* trials in the first place. Thus, on *CX* trials that do not produce an outcome, no error correction takes place, and the absence of the outcome does not have to be explained by an inhibitory influence of *C*.

It might be argued that for the description of conditioned behavior in animals, the failure to accrue inhibitory strength under negative contingencies is of little consequence. Empirically, what is observed is that the animal refrains from executing a particular behavior. For example, if a light signals the absence of an otherwise occurring shock, the animal does not exhibit the typical fear reaction on trials where the light is on. It is difficult to distinguish empirically whether no fear behavior is observed because the light gained inhibitory strength that counters the excitatory strength of the context or whether trials with a light are qualitatively different from background trials with no light and, so, no shock would be expected anyway. When $S_C > S_X$, Pearce's (1987) model endorses the latter explanation.

The failure to accrue inhibitory strength under negative contingencies is clearly problematic when mapping associative learning mechanisms to human causal reasoning, however. The dependent variable is not the presence or absence of a particular behavior; instead we probe participants for estimates of causal strength. If participants interpreted negative contingencies to indicate that contextual causes produce the effect, but expected the effect not to occur on trials when the candidate cause was present (because of lack of generalization between the two), this would show in their causal ratings: They would provide ratings around zero for all preventive conditions. The observed negative ratings in these conditions show that this was not the case: Participants clearly generalized from *X* to *CX* and thus learned that *C* prevents the effect.

One obvious way to allow cues in the Pearce (1987) model to accrue inhibitory strengths under negative contingencies is to change saliency assumptions. For example, assuming $S_C = S_X$ or $S_C < S_X$ shifts the associative strengths of preventive candidates back into the negative space. This is because these assumptions allow for more generalization between *CX* and *X*. However, making these assumptions also reduces the general fit of the model to the ordinal pattern of the data we obtained: The base-rate influence on conditions with identical ΔPs is less pronounced, and, more important, an influence of ΔP on conditions with equal power is erroneously predicted and comparisons between Conditions A and C, and J and F (where ΔP and power make opposite predictions) are in the direction of ΔP, contrary to the observed trend in favor of power.

Implications

Our findings shed light on why purely covariational models, including the associationist RWM, do not adequately describe human causal induction. Lober and Shanks (2000) wrote, "We believe that the kernel of the Rescorla-Wagner model—namely the mechanism of learning via error-correction—is necessary to understand causal learning" (p. 210). But any learning process must necessarily be defined relative to a representation of what is to be learned. A critical element missing from purely covariational models, including associative learning models such as RWM, is an explicit representation of the unobserved distal causal relation. Covariational and associative learning models arrive at representations of causal power by directly transforming covariational evidence into measures of causal or associative strength. The computational causal power approach, on the other hand, postulates that reasoners assume a priori that unobservable causal powers exist in the world around them and that these powers manifest themselves in observable covariations. Causal power in this approach thus is an unbound variable (Holyoak & Hummel, 2000) that may or may not take on specific values. Whether causal power does take a value is determined by boundary conditions (such as absence of ceiling or floor effects; see Wu & Cheng, 1999). Covariational and associative learning models have no such explicit representation. Without representation, an error-correction mechanism cannot operate on the appropriate target of the inductive process, thereby forgoing inference regarding causal relations in the distal world.

If inferring distal causal relations is the goal of human causal induction, however, then the power PC theory (Cheng, 1997)—unlike purely covariational models—is a normative account of this inductive process, and our results show that humans reason about causal relations normatively. This theory provides a computational level description (Marr, 1982) of a solution to the problem of causal induction, of how it is possible for a distal causal relation to be inferred from covariation in the proximal stimuli. Although algorithmic descriptions are desirable complements to computational level descriptions, the appeal of the Rescorla-Wagner algorithm (and similar associative learning algorithms) would be lost if the algorithm cannot yield (or be amended to yield) the output specified at the computational level. Error correction may well turn out to be a central component of the algorithm, but it would likely be in a radically novel form, one that allows the representation of the distal causal relation.[15]

As discussed earlier, a number of robust deviations from rational causal inference have been reported in the literature. These deviations are consistent with past research on human judgment and decision making (e.g., Kahneman, Slovic, & Tversky, 1982) that showed that question format, reference frames, and the type of rating scale used all influence and potentially bias probability and preference ratings. However, rather than resorting to more complex models with additional parameters to accommodate the seemingly robust deviations from rationality, we have chosen to first develop experimental situations that might allow one to understand the causes of the deviations.

The choice between finding good fits to existing data and conducting better controlled experiments can be viewed in a broader context. Any pattern of psychological results is potentially the product of a multitude of distinct psychological processes, many of which are extraneous to the central process in question. These extraneous processes may disguise or mask the output of the process of central concern. If the goal of psychological research is to fit various patterns of data, irrespective of the extraneous psychological processes likely to give rise to those patterns, then models replete with parameters will no doubt have an advantage, as they have additional degrees of freedom. However, fitting arbitrary parameters to fit an unspecified mixture produced by multiple psychological processes will not deepen the understanding of any psychological process.

If the goal of psychological research is instead to understand particular processes that are evidently important, then a more fruitful research strategy is to attempt to isolate the operation of the process under study. Achieving this goal may involve minimizing the influences of extraneous processes that normally would operate under typical everyday situations. The researcher's strategy, just like that of an everyday causal reasoner operating in accord with the power PC theory, is to control for and reduce the influence of alternative causes of the effect in question, so that the candidate causal process can manifest itself as clearly as possible. Thus, contrived situations in which alternative causes are eliminated or controlled can in fact be most informative. What justifies the creation of such situations is the simple and coherent explanation of a complex pattern of findings.

[15] Because causal strength has to be represented as an *unbound* variable (Holyoak & Hummel, 2000), separate from the value of covariation (cf. Wu & Cheng, 1999), any plausible algorithm of human causal induction will have to be of a *symbol manipulating* nature. In contrast, the error-correction mechanisms of the RWM and other similar associative learning models are *symbol eliminating*.

References

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

Baker, A. G., Vallee Tourangeau, F., & Murphy, R. A. (2000). Asymptotic judgment of cause in a relative validity paradigm. *Memory and Cognition, 28*(3), 466–479.

Buehner, M. J., & Cheng, P. W. (1997). Causal induction: The power PC theory versus the Rescorla-Wagner model. In the *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 55–60). Hillsdale, NJ: Erlbaum.

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

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

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments and Computers, 25,* 257–271.

Danks, D. (in press). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology.*

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 229–263). Mahwah, NJ: Erlbaum.

Hume, D. (1902). An enquiry concerning human understanding. In L. A. Selby-Bigge (Ed.), *Hume's enquiries.* Oxford, England: Clarendon Press. (Original work published 1777).

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

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and applied, 79*(1, Whole No. 594).

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

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

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

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: Freeman.

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

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418.

Pearce, J. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94,* 61–73.

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

Perales, J. C., & Shanks, D. R. (in press). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *Quarterly Journal of Experimental Psychology.*

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.

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

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

Shanks, D. R. (1987). Acquisition functions in contingency 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., & 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.

Spellman, B. A. (1996). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation* (Vol. 34, pp. 167–206). San Diego, CA: Academic Press.

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

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Vallee Tourangeau, F., Murphy, R. A., Drew, S., & Baker, A. G. (1998). Judging the importance of constant and variable candidate causes: A test of the power PC theory. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 1,* 65–84.

Van Hamme, L. J., Kao, S., & Wasserman, E. A. (1993). Judging interevent relations: From cause to effect and from effect to cause. *Memory and Cognition, 21,* 802–808.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgment: The role of nonrepresentation of compound stimulus elements. *Learning and Motivation, 25,* 127–151.

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

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

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science, 10,* 92–97.

Received October 15, 2002
Revision received April 9, 2003
Accepted April 14, 2003 ■