Cognitive Science 31 (2007) 765–814

Adaptive Non-Interventional Heuristics for Covariation Detection in Causal Induction: Model Comparison and Rational Analysis

Masasi Hattori[a], Mike Oaksford[b]

[a]*Department of Psychology, Ritsumeikan University, Japan*
[b]*School of Psychology, Birkbeck College London*

Received 3 March 2004; received in revised form 13 January 2007; accepted 29 January 2007

Abstract

In this article, 41 models of covariation detection from 2 × 2 contingency tables were evaluated against past data in the literature and against data from new experiments. A new model was also included based on a limiting case of the normative phi-coefficient under an extreme rarity assumption, which has been shown to be an important factor in covariation detection (McKenzie & Mikkelsen, 2007) and data selection (Hattori, 2002; Oaksford & Chater, 1994, 2003). The results were supportive of the new model. To investigate its explanatory adequacy, a rational analysis using two computer simulations was conducted. These simulations revealed the environmental conditions and the memory restrictions under which the new model best approximates the normative model of covariation detection in these tasks. They thus demonstrated the adaptive rationality of the new model.

Keywords: Causal induction; Contingency judgment; Covariation assessment; Adaptive rationality; Conditional reasoning

1. Introduction

To respond adaptively in an environment that changes from moment to moment, people must make rapid and precise predictions of the future based on current information. In this respect, the concept of causality is central and has probably been acquired over the long course of evolution along with other cognitive abilities (Toda, 1983). There have been many arguments in philosophy and psychology about the concept of causality and the mechanisms of causal judgment (e.g., see, Sosa & Tooley, 1993; Sperber, Premack, & Premack, 1995). However, philosophical discussion about the normative definition of causation, although undoubtedly

Correspondence should be addressed to Masasi Hattori, Department of Psychology, Ritsumeikan University, 56-1 Kitamachi, Toji-in, Kita-ku, Kyoto 603–8577 Japan. E-mail: hat@lt.ritsumei.ac.jp

important, could be irrelevant to descriptive theories of human causal cognition. Such discussions, as they are relevant to science, concern the justification of our causal inferences when under no time pressure and when we have many resources to devote to the problem. However, for survival in the everyday world, for people who have limited memory and limited information processing capacity, it is important to be able to derive plausible conclusions rapidly using minimal resources (i.e., storing as little information as possible).

Most models of causal induction assume that extracting covariation information is an important first step (e.g., Anderson & Sheu, 1995; Baker, Murphy, Vallée-Tourangeau, 1996; Cheng, 1997; Cheng & Novick, 1992; Einhorn & Hogarth, 1986; Schustack & Sternberg, 1981; Shanks & Dickinson, 1987; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996). Of course, however, covariation is not causation. For example, the rooster's crow is not a cause of the sun's rising while they covary. A subsequent analytic process, which may be based on prior domain-specific causal knowledge, is required to identify genuinely causal links between events. Nevertheless, covariation detection is important for us for two reasons. First, there are always innumerable irrelevant factors. We have to exclude a vast number of unrelated events as potential causal candidates to discriminate genuine from spurious causes. Covariation can be a useful cue to screen out unnecessary factors. Second, correlational knowledge is sufficient for predicting the future. As long as we do not intervene, a strong correlation can provide precise predictions. In this article, we concentrate on the covariation-based aspect of causal induction. Our goal is to discover the nature of people's covariation detection mechanism and whether there are circumstances in which it can be given a rational interpretation.

In most experiments, covariation information is represented in a 2×2 contingency table as shown in Table 1. This table shows the relationship between a candidate cause and an effect. Four combinations arise: The candidate cause is present or absent (C or $\bar{C}$, respectively) and the effect is present or absent (E or $\bar{E}$, respectively), leading to the four cells ("C and E," "C and $\bar{E}$," "$\bar{C}$ and E," "$\bar{C}$ and $\bar{E}$"), which are labeled cells a, b, c, and d, respectively. Models of covariation detection can be expressed using the frequencies of these cells.

Experimental studies have revealed that normative models of covariation detection from 2×2 contingency tables are descriptively inadequate. For example, it has been argued that participants do not equally weight the four cell frequencies: the a-cell is the most and the d-cell is the least weighted (e.g., Arkes & Harkness, 1983; Crocker, 1982; Jenkins & Ward, 1965; Kao & Wasserman, 1993; Mandel & Lehman, 1998; Nisbett & Ross, 1980; Schustack & Sternberg, 1981; Shaklee & Mims, 1982; Smedslund, 1963; Ward & Jenkins, 1965; Wasserman, Dorner,

Table 1
A 2×2 contingency table containing covariation information

	Effect	
Cause	Present (E)	Absent ($\bar{E}$)
Present (C)	a	b
Absent ($\bar{C}$)	c	d

Note: In areas other than causal induction, combinations of a cause and an effect could be prediction–actuality (e.g., weather forecasting), stimulus–response (e.g., recognition memory, perceptual response), etc.

& Kao, 1990). However, normative indices (e.g., the phi-coefficient [ϕ], introduced later) require all the cells to be weighted equally. This unequal weighting may be a reflection of a psychological asymmetry between the concepts of presence and absence, which is closely related to the fact that presence is often rare while absence is pervasive.

Recently, however, a number of theorists have argued for the adaptive rationality of differentially weighting the cell frequencies (e.g., Anderson & Sheu, 1995; Cheng, 1997; Friedrich, 1993; Gigerenzer & Hoffrage, 1999; Mandel & Lehman, 1998; McKenzie, 1994; but see also, Over & Green, 2001). Although a particular heuristic strategy may seem irrational from a normative standpoint, it may work practically very well given the structure of our natural environment and so can be justified from an adaptive point of view (e.g., see, Anderson, 1990; Evans & Over, 1996a; Gigerenzer, 2000; Oaksford & Chater, 1998a, 1998b; Payne, Bettman, & Johnson, 1993; Stanovich, 1999). McKenzie (1994) investigated the average accuracy of some non-normative indices proposed as descriptive models of covariation judgments from 2×2 contingency tables.[1] He compared them with a normative statistic (the phi-coefficient, ϕ) and showed that many of them performed quite well. His seminal work revealed that some heuristics could be practically very useful, but it is still unclear which model predicts human performance best and, moreover, why people use a particular heuristic.

2. Scope and overview

We propose that human causal induction has two stages: a *heuristic stage* and an *analytic stage*. The analytic stage is essential to discriminate between genuine and spurious causes. The heuristic stage is also vital in distinguishing relevant causal candidates from innumerable irrelevant factors. We believe that different models are needed to explain the complex process of human causal induction and this has not been explicitly recognized. The dual process approach in causal induction is relevant to several two-system views in human cognition (e.g., Evans, 1989; Sloman, 1996; Stanovich, 1999). This study addresses only the issue of the heuristic stage of causal induction and leaves open the question of what else is required to identify genuinely causal relationships in the analytic stage. The heuristic stage of causal induction is mostly concerned with covariation assessment between two events. As a first step in causal induction, covariation assessment is often based on observation, whereas intervention (i.e., experimental manipulation of particular factors) to idenfity real from spurious causes is part of the analytic stage (e.g., Lagnado & Sloman, 2004).

We first introduce a new heuristic index of covariation detection, the *dual factor heuristic* (Hattori, 2001), which is motivated by considering the rarity of particular causes and effects in the environment. This factor has proved important in other related areas (Hattori, 2002; McKenzie & Mikkelsen, 2000, 2007; Oaksford & Chater, 1994, 2003). Next we report the results of an experiment aimed to examine the difference between two common experimental paradigms (discrete vs. continuous) in causal induction. The experiment was intended to decide which data in the literature could be included in meta-analyses to compare the dual factor heuristic with other indices of covariation. We then present a meta-analysis exhaustively comparing this index with 33 other non-parameterized indices that have been proposed in various literatures with respect to their ability to account for the data from a signature set

of past causal induction and covariation detection experiments. We note that the dual factor heuristic comes out best in this comparison but that other indices are very close runners up. We therefore report the results of an experiment designed to discriminate between these indices to determine which is the best of this bunch. We then present a further meta-analysis comparing the dual factor heuristic with parameterized models of covariation detection. Finally, we present a rational analysis of the dual factor heuristic in computer simulations.

The fundamental question that we address is why does the dual factor heuristic fit the data? The answer we will suggest is that under certain reasonable environmental conditions it provides a good approximation to normative predictions. To clarify this point, we conducted two computer simulations using the Monte Carlo method. In the simulations, following Anderson's (1990) *rational analysis*, we take three factors into consideration: the structure of the environment, cognitive constraints (time, memory, load, etc.), and the results of the behavior (accuracy, satisfaction, etc.). It turns out that the factors that determine when the dual factor heuristic best approximates normative predictions in covariation detection are the same as those found in data selection (Hattori, 2002; Oaksford & Chater, 1994, 2003). We first introduce the dual factor heuristic.

3. The dual factor heuristic

3.1. Two stages in causal induction

We have proposed that there is a heuristic and an analytic stage to human causal judgment. These two stages are differently motivated and the dual factor heuristic is concerned with the heuristic stage. Identifying genuine causes is important for us to control the environment and to change the world to meet our needs. Whereas correlation enables us to predict the future, causation provides us with more: it enables us to imagine the results of our own actions on the environment. The analytic stage, however, requires more cognitive resources and time. Given the cognitive costs of the analytic process, deciding which factors should be marked out as potential causal candidates and which ones can be ignored is important for survival in the real world. This *two stage hypothesis* in causal induction parallels *dual process theory* (Evans, 1989) in reasoning.

This hypothesis is also relevant to recent discussion of the distinction between *observation* and *intervention* in causal reasoning (Pearl, 2000). In principle, it is hard to deduce a causal dependency only from observational (non-interventional) data like that presented in 2×2 contiungency tables. Correlation does not imply causation. Intervening between events that covary is essential to learning whether one of these events is the cause of the other. For example, we must prevent the rooster from crowing (i.e., intervene on a causal candidate) to prove that it's crowing is not the cause of the sun rising. Using Pearl's terminology of the "do" operater, we need to $do(\neg Crow)$ to ensure that $\neg Sunrise$ (where "$\neg$" = not). Recently, some researchers have suggested that the cognitive processes involved in intervention differ from those of observation (Lagnado & Sloman, 2004; Sloman & Lagnado, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blumd, 2003). Thus, there is a clear distinction between situations where the presence or absence of the cause and effect is simply observed, and situations where

someone chooses to allow or prevent the cause from occurring and keeps track of the effect. However, before one can get to the intervention strategy one must have some candidate causes in mind and these must derive from covariation detection.

3.2. A heuristic ignoring nonoccurrence

As we pointed out above, normative theories of covariation detection include information about all four possible events relating cause and effect, i.e., information from all four cells (a, b, c, and d) in a contingency table. However, presenting such information in summary form in a contingency table is highly unrealistic. People normally encounter the relevant events that provide the data for covariation detection sequentially and compute covariation online from this stream of data. This involves storing and retaining information about the relevant events. With respect to any particular cause–effect relation, this creates a serious problem.

Although it is clear what should be stored when the cause alone occurs (cell b), the effect alone occurs (cell c) or both cause and effect occur (cell a), it is not clear what information to record when neither cause nor effect occur (cell d). Most of the time, any particular cause and effect are not occurring as we move around the world (just think of any particular event, such as starting your car). How then are these non-occurring events to be counted? Perhaps the time in between events of type a–c could be divided into temporal bins in which the cause or effect could occur and these are counted or stored away as instances of d-cell events. Storage of such events would appear very inefficient and creates a problem analogous to the frame problem in artificial intelligence (McCarthy & Hayes, 1969; Pylyshyn, 1987). The frame problem arose in the context of reasoning about change: It relates to the need to store all the information about what does *not* change in a situation when something else does (Oaksford & Chater, 1991, 1993, 1995, 1998b; also explored the ramifications of this problem for theories of human deductive reasoning). So, if the coffee was spilt, then frame axioms would have to be added to the database indicating that the picture did not fall to the floor, the window did not open, etc. Storage of such information is computationally prohibitive as the information to be stored soon exceeds reasonable bounds. This problem also arises in the context of covariation detection, all potential cause-effect relations will have to have non-occurring d-cell events stored as their respective d-cell instances.

A more efficient procedure would be to store information about a–c cell events and infer the d-cell. As for any particular cause–effect relation the cause and effect are not normally present, the inference is relatively clear cut: The d-cell is likely to be very large. The assumption that the d-cell is very large is equivalent to the assumption that the base rates of the cause, $P(C)$, and the effect, $P(E)$, are small. This is consistent with the rarity assumption that has proved important in providing rational analyses of data selection behavior (Evans & Over, 1996b; Klauer, 1999; Nickerson, 1996; Oaksford & Chater, 1994, 2003) and in providing a Bayesian justification for differential cell weightings in judging covariation (e.g., Anderson & Sheu, 1995; McKenzie & Mikkelsen, 2007).

In deriving an index of the dual factor heuristic for non-interventional covariation detection, we consider a normative statistical index, the *phi-coefficient*, ϕ, because it is the most popular measure of correlation. We explored the consequences of assuming the d-cell is very large

for ϕ (we thank an anonymous reviewer for suggesting this derivation). In terms of the cell frequencies this coefficient can be rewritten as follows (from Formula 22 in Table 2):

$$\phi = \frac{a}{\sqrt{(a+b)(a+c)I}} - \frac{bc}{d\sqrt{(a+b)(a+c)I}}. \tag{1}$$

Here, $I = 1 + \frac{b+c}{d} + \frac{bc}{d^2}$. Taking the limit as $d \to \infty$, $I \to 1$ and the second term of Equation 1 goes to zero, as every term containing d or d^2 in the denominator approaches zero. Consequently, we have:

$$\lim_{d\to\infty} \phi = \frac{a}{\sqrt{(a+b)(a+c)}} = \sqrt{P(E|C)P(C|E)} \triangleq H. \tag{2}$$

We refer to this index as the *dual factor heuristic*, H, because it is equivalent to the geometric mean of two factors: the probability of the cause given the effect, $P(C|E)$, and the probability of the effect given the cause, $P(E|C)$. Rewriting ϕ in terms of probabilities as follows can yield an insight into the nature of H:

$$\phi = \sqrt{\frac{P(E|C) - P(E)}{1 - P(E)} \cdot \frac{P(C|E) - P(C)}{1 - P(C)}}. \tag{3}$$

Comparing Equations 2 and 3, we can see that H ignores the base rates of the cause and the effect, $P(C)$ and $P(E)$. Although ϕ has a high value only when $P(E|C)$ and $P(C|E)$ are high by comparison with the corresponding base rates, high values of $P(E|C)$ and $P(C|E)$ per se raise H. This is also obvious because when $d \to \infty$, $P(C), P(E) \to 0$.

The argument that leads to the dual factor heuristic, in short, is as follows. Normative assessments of the strength of covariation require attending to all events, including those that involve nothing happening (d-cell). Attending to all events places prohibitive demands on our limited memory capacity. Computing H allows you to assess the strength of relations without using the d-cell. This is more efficient on memory and is accurate under circumstances where d is large—that is, $P(E)$ and $P(C)$ are small (i.e., the situation people normally encounter in the real world). Later on we also see that H is actually even better than using the d-cell if one is only allowed to keep track of a small number of events.

*3.3. What "*d*-cell insensitivity" is not*

We have argued that the dual factor heuristic, that ignores the d-cell, can be an efficient method for covariation detection in "normal" situations. This claim, however, could be subject to misunderstanding. Consquently, here we seek to clarify what d-cell insensitivity means by clarifying what it is not.

First, d-cell insensitivity is not a universal law of human causal induction. According to our account the dual factor heuristic operates at the first heuristic stage of causal induction, picking out possible causal candidates. In the subsequent analytic stage, people identify genuine causes from this candidate set perhaps using interventional strategies. Interventional strategies, where people actively make the cause happen or prevent it from happening, should place as much emphasis on causal necessity—for example, if $do(\neg Crow)$ then $\neg Sunrise$, as

Table 2
Probabilistic Models for 2 × 2 Contingency Tables

No.	Model/Index/Measure	Definition
1	Dual-factor hueristic	$H \triangleq a/\sqrt{(a+b)(a+c)}$
2	ΔP rule	$\Delta P \triangleq (ad-bc)/[(a+b)(c+d)]$
3		$\Delta P^c \triangleq (ad-bc)/[(a+c)(b+d)]$
4	Power PC (Cheng, 1997)	$PW \triangleq (ad-bc)/(a+b)d$
5		$PW^c \triangleq (ad-bc)/(a+c)d$
6	Probability of necessity (Pearl, 2000)	$PN \triangleq (ad-bc)/a(c+d)$
7		$PN^c \triangleq (ad-bc)/a(b+d)$
8	Causal support (Griffiths & Tenenbaum, 2005)	$SP \triangleq \log[P(D\|G_C)/P(D\|G_I)]$
9	SDT-based measure	$SDT \triangleq \Phi[\Phi^{-1}[d/(c+d)] - \Phi^{-1}[b/(a+b)]]$
10		$SDT^c \triangleq \Phi[\Phi^{-1}[d/(b+d)] - \Phi^{-1}[c/(a+c)]]$
11	Inhelder and Piaget (1958)	$R \triangleq [(a+d)-(b+c)]/N$
12	Proportion correct 1	$C \triangleq a/(a+b)$
13		$C^c \triangleq a/(a+c)$
14	Proportion correct 2	$PC \triangleq (a+d)/N$
15	Positive test (Klayman &Ha, 1987)	$PT \triangleq a(2a+b+c)/2(a+b)(a+c)$
16	Positive hits	$L_1 \triangleq a$
17	Hits minus false positives	$L_2 \triangleq a-b$
18		$L_2^c \triangleq a-c$
19	Sum of diagonals (ΔD)	$L_3 \triangleq (a+d)-(b+c)$
20	Aggregate model (McKenzie, 1994)	$L_4 \triangleq 4a-3b-2c+d$
21		$L_4^c \triangleq 4a-2b-3c+d$
22	Four-fold point correlation coefficient	$\phi \triangleq (ad-bc)/\sqrt{(a+b)(c+d)(a+c)(b+d)}$
23	Chi-square statistic	$\chi^2 \triangleq N(ad-bc)^2/[(a+b)(c+d)(a+c)(b+d)]$
24	Yule's (1900) coefficient of association	$Q \triangleq (ad-bc)/(ad+bc)$
25	Yule's (1912) coefficient of colligation	$Y \triangleq (\sqrt{ad}-\sqrt{bc})/(\sqrt{ad}+\sqrt{bc})$
26	(Goodman and Kruskal, 1954, 1963)	$\lambda_{q\|p} \triangleq \frac{\max(a,b)+\max(c,d)-\max(a+c,b+d)}{\min(a+c,b+d)}$
27		$\lambda_{p\|q} \triangleq \frac{\max(a,c)+\max(b,d)-\max(a+b,c+d)}{\min(a+b,c+d)}$
28		$\lambda \triangleq \frac{\max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)-\max(a+b,c+d)-\max(a+c,b+d)}{\min(a+c,b+d)+\min(a+b,c+d)}$
29	Kappa statistic (Cohen, 1960)	$\kappa \triangleq 2(ad-bc)/[(a+b)(b+d)+(a+c)(c+d)]$
30	Skill test	$S_k \triangleq 4(ad-bc)/N^2$
31	A measure based on Good (1961)	$G \triangleq (ad-bc)/[d(a+b)+b(c+d)]$
32		$G^c \triangleq (ad-bc)/[d(a+c)+c(b+d)]$
33	A measure based on Suppes (1970)	$S \triangleq a/(a+b)-(a+c)/N$
34		$S^c \triangleq a/(a+c)-(a+b)/N$

(*Continued on next page*)

Table 2
Probabilistic Models for 2 × 2 Contingency Tables (*Continued*)

No.	Model/Index/Measure	Definition
35	Asymptotic model of Pearce (1987)	$J \stackrel{\triangle}{=} \frac{x_2[a(c+d)-c(x_1a+x_1b)]}{a(c+d)+b(c+d)-x_1c(x_1a+x_1b)-x_1d(x_1a+x_1b)}$
36	Weighted ΔP rule 1	$\Delta P_{w1} \stackrel{\triangle}{=} a/(a + w_1b) - c/(c + w_2d)$
37	Weighted ΔP rule 2	$\Delta P_{w2} \stackrel{\triangle}{=} \beta_0 + \beta_1 a/(a + b) - \beta_2 c/(c + d)$
38	Information integration model 1	$I_1 \stackrel{\triangle}{=} w_0 + (a - w_1b - w_2c + w_3d)/(a + w_1b + w_2c + w_3d)$
39	Information integration model 2	$I_2 \stackrel{\triangle}{=} \beta_0 + \beta_1 a/N - \beta_2 b/N - \beta_3 c/N$
40	Simple Bayesian (Anderson & Sheu, 1995)	$B \stackrel{\triangle}{=} 1 - 1/[1 + \frac{p_r {p_c}^a (1-p_c)^b {p_e}^c (1-p_e)^d}{(1-p_r){p_n}^{a+c}(1-p_n)^{b+d}}]$
41	Weighted linear combination model	$L_* \stackrel{\triangle}{=} \beta_0 + \beta_1 a + \beta_2 b + \beta_3 c + \beta_4 d$

Note: In the Definition column, a, b, c, and d indicate frequencies of the events "C and E," "C and $\bar{E}$," "$\bar{C}$ and E," and "$\bar{C}$ and $\bar{E}$," respectively (see Table 2); and $N = a + b + c + d$.
The right hand superscript c indicates that the index is a *converse index* of the original one, namely, the index calculated by interchanging the rows and columns of the 2 × 2 table.

SP [No. 8] (Griffiths & Tenenbaum, 2005) cannot simply be described using cell frequencies. See Appendix A for detail.

sufficiency, if *do*(*Crow*) then *Sunrise*. This emphasis on causal necessity in intervention, *ipso facto* requires taking the d-cell into account because causal necessity is high when $P(\neg Crow, \neg Sunrize)$ is high. Thus, one would expect a simple heuristic strategy like H to be much less in evidence in paradigms where interventional strategies are used. In Experiment 1, we contrast the discrete and continuous paradigms investigated by Anderson and Sheu (1995). We argue that the discrete paradigm is mainly observational, whereas the continuous paradigm is more interventional.

Moreover, weak d-cell effects are compatible with a small subgroup of participants using such an analytic strategy even with observational data. This is similar to the case of occasional logical performance in deductive reasoning tasks. Oaksford and Chater's (1994, 2003) probabilistic approach explains this phenomena as follows. Whereas most participants use heuristic or low level probabilistic strategies, those with high IQs have learned logical strategies that they can apply to the tasks (Stanovich, 1999). This explains occasional logical performance and why low level probabilistic strategies provide much better fits to the data than logical strategies. By analogy, a few participants may be able to apply higher level strategies even to purely observational data just as in deductive reasoning some participants can use logical strategies. This is one of the reasons why we sometimes observe weak d-cell effects and why for observational data we would expect to find better fits for indices like H that ignore the d-cell.

Second, d-cell insensitivity is not a norm or a golden rule for causal induction. The dual factor heuristic is an approximation that is effective in the real world considering limited cognitive resources. If we need to assess a correlation between two events *ad unguem* (perhaps in the analytic stage) and we have enough time and resources to do so, we should use a normative index such as phi-coefficient or ΔP, which exploits the d-cell.

Third, d-cell insensitivity is not necessarily a "null coefficient" for the d-cell in a certain parameterized model, particularly in a multiple linear regression model, which is known as *weighted linear combination model* for covariation detection (see Formula 41 in Table 2 and Equation 11, introduced later). Without definite evidence that a regression model best describes people's covariation detection, fitting the model and judging the importance of a variable (i.e., the effect of corresponding cell) based on the magnitude of (standardized) regression coefficients can lead to erroneous conclusions. For example, consider a set of data, $(x, y, z) = (2, 1, 4), (3, 3, 9), (4, 6, 16)$, where x and y are predictors and z is the dependent variable. These data are completely fitted by a simple non-linear model, $z = x^2$, which "ignores" y (i.e., it is assigned a zero weight). The same data set, however, are also fitted by a linear model, $z = x + 2y$, which has non-null coefficient (i.e., 2) for y. As shown in this example, it can be a case that a set of data generated by a model in which a certain variable (e.g., d-cell) is missing is also well fitted by a linear regression model with a non-null coefficient for the variable. Thus, just because the d-cell is assigned a non-null weight in a multiple regression does not mean that a model that uses the d-cell is the best model for that data. In particular, a model that uses less information to achieve the same level of fit would be preferred as more parsimonious. The long-standing observation of weak d-cell effects in covariation detection (or causal induction) is mostly based on linear regression models (e.g., Schustack & Sternberg, 1981). However, one could only determine whether a model that uses the d-cell is required by a comparative model fitting exercise like those we report in this article.

Consequently, despite the empirical findings on the d-cell effects, the question still remains as to whether an index of covariation that ignores the d-cell such as H can provide better accounts of the evidence than alternative indices.

3.4. Preventive causes

The case of a preventive cause might be seen as just a mirror image of the case of a generative cause such that only the truth value of the effect is reversed, so that "C prevents E" is just to say that "C causes $\bar{E}$." If this were the case, then preventive causes are also formalized in our framework—reversing the truth value of the effect results in swapping the a-cell with the b-cell and the c-cell with the d-cell.[2] Formally, the dual factor heuristic index for preventive causes could then be expressed as $b/\sqrt{(a+b)(b+d)}$.

However, there seems to be more to the concept of prevention. First, note that flipping the truth value changes the probability associated with the event: For example, when $P(E)$ is small, $P(\bar{E})$ is large. According to the dual factor heuristic, when the sets C and E almost totally overlap, the predicted strength of covariation is large. However, when C is small and E is large this strength decreases. Therefore, for H to take on a high value for a putative preventive cause, when $P(C)$ is small, $P(\bar{E})$ should also be small and hence $P(E)$ should be large (e.g., "committing a felony [C] *prevents* being at work [E]"). As the dual factor heuristic assumes the rarity of the cause and the effect, in the case of preventive causes, prevalence of the effect should be maintained.

In addition to reversing the truth value, most statements of preventive causation require *re-contextualization*, or recircumscribing the causal field (similar processes are also described by, e.g., Cheng, 1997). For example, the statement "ingesting vitamin C prevents catching a

cold" does *not* exactly mean "ingesting vitamin C causes not-catching a cold" because even if we do not ingest vitamin C, we do not usually catch a cold. The intended class of events is not general situations in our daily life but those where some other conditions are met for catching a cold, for example, when your partner has a cold. Consequently, paraphrasing "C prevents E" as "C causes $\bar{E}$" requires refocusing or recontextualizing on the appropriate causal field. Without doing so the d-cell can extend unreasonably and this can lead to anomalies even for normative indices.

Let us look at another example, "the sarin (poison gas) antidote prevents sarin poisoning." First we consider the case without appropriate contextualization to "sarin exposure" events. Reversing the truth value of the effect and replacing "cause" with "prevent," we have to take into account the vast number of d-cell cases (i.e., no-antidote and no-poisoning), which contributes to positive diagnoses for some (normative) indices. Suppose here that the observed frequencies of an un-contextualized sample were 1, 4, 5, and 10^6 for the cells a (antidote and poisoning), b (antidote and non-poisoning), c (non-antidote and poisoning), and d (non-antidote and non-poisoning), respectively—almost all people did not have sarin antidote and did not experience sarin poisoning. In this case, $\phi = .18$ and $\Delta P = .20$ (i.e., both indicate a weak positive correlation)—that is, counterintuitively, rather than preventing sarin poisoning these indices indicate that the antidote is a weak cause of sarin poisoning!

We now consider the appropriately contextualized case where the relevant class of events is sarin exposures. If all the d-cell observations actually came from sarin non-exposure cases, the "real" d-cell frequency is 0 and $\phi = -.82$ and $\Delta P = -.80$, which are both strongly negative, as expected. After appropriate recontextualization, we can treat the target case as a mirror image of a generative cause (i.e., we can paraphrase " . . . *prevents* sarin poisoning," as " . . . *causes* protection against sarin poisoning"). Consequently, cell a can be swapped with b and c with d in the contingency table. For this table, $H = .89$ (strong negative, as expected). Thus, the dual factor heuristic can handle the case of a preventive cause after appropriate re-contextualization, rephrasing, and swapping the columns of the contingency table.

Recontextualization and reversing the truth value might be viewed as too complex for a fast heuristic for covariation detection. However, in text comprehension, for example, it is known that we automatically and rapidly produce online inferences about the contextually appropriate meanings of words (on the order of few hundred milliseconds) to establish local coherence (e.g., McKoon & Ratcliff, 1992). Likewise, encoding the polarity of the effects can be achieved by a simple "take the smaller" rule—that is, people compare the set sizes of the target and its complement (negative) proposition in the context (e.g., "being employed" and "being unemployed") and take the smaller (i.e., rarer) one (i.e., "being unemployed") to detect the covariation. In expressing causal relations, focusing on the rare case is exactly what people do (McKenzie, Ferreira, Mikkelsen, McDermott, & Skrable, 2001).

Although some models of causal induction (e.g., ΔP) deal with generative and preventive causes within a single framework the dual factor heuristic basically provides a method to detect generative causes either between $C$and E or between C and $\bar{E}$ dependent on which is rarer E or $\bar{E}$. This might be regarded as a disadvantage of this model. However, the function of the heuristic stage, as in other two stage models (Evans, 1989), is *relevance* detection (i.e., finding out whether a potential cause relevant enough to be considered a candidate for subsequent testing at the analytic stage). It is only at the analytic stage that genuine causes

are separated from spurious causes and the precise nature of the causal relation, preventive or generative, are fully determined by intervention in the causal system itself rather than by simply observing it.

In the following, we first determined whether the dual factor heuristic provides a good account of covariation detection by comparing its performance to all other indices on 2×2 contingency tables that we could find in the literature. We now introduce these other indices before turning to a comparison based on a meta-analysis of some of the past data on causal induction and covariation detection.

4. Indices for 2 × 2 contingency tables

There have been many proposals for appropriate measures of the strength of covariation. However, there have been few studies that have exhaustively compared these models with each other. There seem to be two reasons for this. First, instead of comparing models for how well they fit the data, some studies (e.g., Anderson & Sheu, 1995; Schustack & Sternberg, 1981) are devoted to calibrating a specific model with parameters. Second, some other studies (e.g., Shanks, Lopez, Darby, & Dickinson, 1996; Wasserman et al., 1996) have concentrated on specific features of causal induction such as *learning effects* (or *sample size effects*) or *overshadowing* (or *discounting*), which are relevant only to the disagreement between proponents of associative learning models (e.g., Rescorla & Wagner, 1972) and proponents of statistical contingency models, particularly, the so-called ΔP rule (e.g., Cheng & Novick, 1992). Consequently, a subsidiary goal, in addition to determining the adequacy of the dual factor heuristic, was to exhaustively investigate the descriptive validity of various models of causal judgment from 2×2 contingency tables.

A wide variety of relations can be described by a 2×2 contingency table other than covariation including, "prediction and accuracy" (e.g., in studying the feeling-of-knowing and in weather forecasting) and "stimulus and response" relations (e.g., recognition memory and perceptual responses). Although each relational concept differs in background motivations, many of these concepts overlap. They can be roughly placed in the following two categories according to how they are motivated:

1. Normative studies of measures in psychology, statistics, epidemiology, medical science, and meteorology. The relations studied have been labeled "accuracy," "agreement," "association," or "correlation."
2. Psychological descriptive studies of human or animal judgment or learning. The relations studied have variously been labeled "causation," "contingency," "correlation," or "covariation."

Some articles provide detailed reviews of these indices within a certain restricted domain (e.g., Allan, 1980; Brier & Allen, 1951; A. W. F. Edwards, 1963; J. H. Edwards, 1957; Goodman & Kruskal, 1954; Landis & Koch, 1975a, 1975b; Nelson, 1984; Swets, 1986; Woodcock, 1976). Below, ignoring differences of backgrounds, we comprehensively review these indices. Table 2 shows the complete list of the indices investigated in this study. In the

text below, a number in brackets followed by an index name or its description corresponds to the one in this table.

Before reviewing indices, some points should be noted. First, some indices are symmetrical, whereas others are not. Namely, if we interchange the cause and effect by swapping rows and columns of the contingency table, some indices show a different value. We also test these "converse" indices, which are indicated by the superscript c in Table 2.

Second, as we mentioned above, some models developed to describe human and animal behavior have adjustable parameters including the *weighted* ΔP *model* (e.g., Allan, 1993; Anderson & Sheu, 1995), the *weighted linear combination model* (e.g., Schustack & Sternberg, 1981), and the Rescorla–Wagner model (Rescorla & Wagner, 1972). In these models, parameters are introduced and estimated so that the model fits the data; and, as a result, estimated parameter values vary from data set to data set. We first compared H to non-parameterized models in a meta-analysis of previous data. We then compared H to parameterized models using some of this data. This was because different measures of fit are required in each case and not all the data is sufficiently rich to constrain the parameters of the parameterized models.

The third issue is concerned with the ranges of the various indices. Some indices run over infinite ranges (e.g., from 0 to plus infinity $[+\infty]$), whereas others have limited ranges (e.g., from 0–1). Moreover, some indices (e.g., χ^2) are sensitive to sample sizes, wheras the others (e.g., ϕ) are not. On the one hand, theoretically speaking, the sample size can be increased without limit; it would thus be reasonable to assume that the indices sensitive to sample size do not have an upper limit. On the other hand, as for indices that have an infinite range despite insensitivity to sample size, we should standardize their ranges using a proper monotonic function to make a fair competition. This is because our analyses here are based on linear correlations between each model and participants' causal ratings with a limited range (e.g., from 0–100), although we back up the results with non-parametric method. Each method of standardization was chosen so as to be suitable to each formula, which is detailed in Appendix A.

4.1. ΔP rule and related indices

Among the many models of causal induction that have been developed based on covariation, the ΔP *rule* [2, 3] is a well-known standard:

$$\Delta P \stackrel{\Delta}{=} P(E|C) - P(E|\bar{C}) = \frac{ad - bc}{(a + b)(c + d)}. \tag{4}$$

The ΔP rule has been proposed to explain human judgment of causality or correlation (e.g., Allan, 1993; Allan & Jenkins, 1980; Cheng & Novick, 1992; Jenkins & Ward, 1965; Ward & Jenkins, 1965) or as an index of association between the conditioned stimulus (CS) and the unconditioned stimulus (US) in Pavlovian learning in animals (e.g., Rescorla, 1968). Recently, it was shown that the Rescorla–Wagner model (Rescorla & Wagner, 1972) of associative learning converges to a *weighted* version of ΔP, introduced later (Wasserman, Elek, Chatlosh, & Baker, 1993), or simply to ΔP (Chapman & Robbins, 1990; Cheng, 1997) under certain constraints (see, Danks, 2003, for details). This index has also been the subject of much interdisciplinary discussion in the area of causal cognition (e.g., Cheng & Holyoak, 1995; Holland, Holyoak, Nisbett, & Thagard, 1986; Sperber et al., 1995).

It is, however, perhaps less well-known that accuracy has been measured by the same index. ΔP is the equivalent to an index that has been used to describe the accuracy of recognition memory in psychology (Gillund & Shiffrin, 1984; Woodworth, 1938) or the accuracy of prospects in meteorology and medical diagnosis (Hanssen & Kuipers, 1965, described in Woodcock, 1976). For example, in the case of recognition memory, the index indicates the difference of proportions between correct recognition of the old items and incorrect recognition of the new items. Moreover, an index known as the *Hart difference score* (Hart, 1965), which has been used in studies of meta-cognition as a measure of accuracy of feeling-of-knowing (Nelson, 1984), is also equivalent to ΔP.

The *power PC theory* (Cheng, 1997) is a modification of the *probabilistic contrast model* (Cheng & Novick, 1990, 1992), which is based on the ΔP rule and the concept of a *focal set*. According to this theory, when it is not known that there is another cause that is dependent on the candidate cause, *causal power* (*PW*) [4,5] is defined as:

$$PW \stackrel{\Delta}{=} \frac{\Delta P}{1 - P(E|\bar{C})} = \frac{ad - bc}{(a + b)d}. \tag{5}$$

Interestingly, this index is formally equivalent to an index that has been used in studies of perceptual function (Blackwell, 1963; Fisk & Schneider, 1984), although the background theory is conceptually different. With respect to visual discrimination tasks, Blackwell regarded this index as the ideal probability of "yes" responses in a sensory discrimination task, provided that $P(E|\bar{C})$ is the probability of giving (spurious) "yes" responses when sensory discrimination is absent, and $P(E|C)$ is the total probability of "yes" responses due both to sensory discrimination and spurious "yes" responses. This idea is connected with the following approaches based on graphical models of causality.

Pearl (2000) constructed a new system to express and compute causality. He defined three indices that measure causal strength (chap. 9): the probability of necessity (*PN*) [6, 7], the probability of sufficiency (*PS*), and the probability of necessity and sufficiency (*PNS*). According to his definitions, in the simplest situation[3] $PN = \Delta P \,/\, P(E|C)$, *PS* coincides with *PW*, and *PNS* coincides with ΔP.

Griffiths and Tenenbaum (2005; Tenenbaum & Griffiths, 2001) showed that both ΔP and *PW* can be regarded as maximum likelihood estimates of causal strength in a graphical model, G_C, where three variables, an effect (E), a potential cause (C), and a set of background factors (B), are involved. In G_C, there are two directed links connected to E: $C \xrightarrow{p_C} E$ and $B \xrightarrow{p_B} E$. In case C occurs, E occurs with probability p_C. Likewise, B produces E with probability p_B. The causal strength is defined as a probability p_C. ΔP and *PW* only differ in functional relations (parameterizations) between the cause and effect. Treating causal induction as causal structure learning, which can be formalized as a decision between two causal graphical models, G_C and G_I, Griffiths and Tenenbaum proposed a new model called *causal support*, *SP* [8]. G_I is a model for causal irrelevance in which there is only a link between C and B. Causal support is defined by the value of the log likelihood ratio for obtaining data D from G_C over G_I as follows:

$$SP \stackrel{\Delta}{=} \log \frac{P(D|G_C)}{P(D|G_I)}. \tag{6}$$

$P(D|G_C)$ and $P(D|G_I)$ are defined by integrating over all possible values that p_C and p_B could assume as shown in Appendix A. *SP* cannot simply be expressed using cell frequencies.

4.2. Other normative models in psychology

Signal Detection Theory (SDT; Green & Swets, 1966/1988; Tanner & Swets, 1954) [9, 10] is based on an analogy between the way the mind works in sensory discrimination or detection tasks and the Neyman–Pearson theory of statistical hypothesis testing (Gigerenzer & Murray, 1987). It proposes two internal distributions: "noise" alone and "signal and noise." Each of these distributions corresponds to distributions of some sample statistic (e.g., mean) of two competing hypotheses in Neyman–Pearson theory. *Detectability*, d', is calculated using two probabilities: $P(E|\bar{C})$—that is, the probability of a "yes" response under noise alone (*false alarms*); and $P(E|C)$—that is, the probability of a "yes" response under signal plus noise (*hits*; see Appendix A for more detail).

Likewise, *Choice Theory* (Luce, 1959, 1963) relates to human decision processes. Luce (1963) defined a measure of similarity, η, between signal and noise and argued that –log η could be regarded as *mental distance* (p. 113), which is similar to d'. Although these two indices may be closely related to human causal judgment, they both need to be standardized because they run from $-\infty$ to $+\infty$. It turns out that using a standard transformation, an index based on Choice Theory is formally equivalent to Yule's (1912) *coefficient of colligation* [25] (introduced below), so this index was omitted (see Appendix A for the method).

Inhelder and Piaget (1958) formalized the concept of correlation as the difference between $(a + d)$ and $(b + c)$ to investigate children's conceptual development. They proposed the *association coefficient* [11] as the simplest measure of correlation (p. 234). This index is a variation of a well known strategy called the *sum of diagonals* or ΔD (introduced later as L_3 [19]) standardized by the sample size. More recently, White (2003) has shown that some people use this strategy which he calls *pCI*.

The *proportion correct* has been used frequently as a measure of response accuracy because of its simplicity and intuitive appeal. There are two ways to define the proportion correct: pay attention only to positives (i.e., the cause is present) [12,13], C, or including negatives [14], PC. PC was first used more than 100 years ago to evaluate the accuracy of predicting tornadoes in America (e.g., Finley, 1884, described in Nelson, 1984) and has also been used as a measure of causal strength (White, 2000) or correlation (Smedslund, 1963). Although these indices are often used in psychology, meteorology, and medical diagnosis, they are margin sensitive and considered inappropriate as normative measures of accuracy (for more argument, see, e.g., Nelson, 1984; Swets, 1986).

4.3. Normative models in statistics

Probably the most widely used statistical measure of correlation between two discrete variables is the *four-fold point correlation coefficient* [22], usually known as *phi-coefficient*, ϕ. It relates to two binary variables, and it corresponds to *Pearson's product–moment correlation coefficient*, r, which relates to two continuous variables. In addition to ϕ, there are two other

standard indices of correlation in a 2×2 contingency table: Yule's (1900) *coefficient of association* [24], Q; and Yule's (1912) *coefficient of colligation* [25], Y (see Appendix A).

The chi-square statistic [23], χ^2, is also a well-known measure of the relation between two discrete variables. It is strongly related to ϕ as can be seen from their definitions (see Table 2). Contrary to ϕ, χ^2 is sensitive to sample size. This is because χ^2 is concerned with statistical judgments of the dependency between two discrete variables, which is affected by both correlational strength and sample size.

Goodman and Kruskal (1954, 1963) defined two asymmetric statistics [26,27] both based on the principle of predicting one variable from another, which is known as *proportional reduction in error* (PRE; e.g., see, Bishop, Fienberg, & Holland, 1975). Suppose one needs to predict the category of a discrete variable E in two different situations: (a) where nothing else is known and (b) where information about C is known. Generally, the probability of error in b is less than in a. Thus PRE is defined as the relative degree of improvement in terms of prediction of E given C:

$$\lambda_{E|C} = \frac{\text{Prob. of error in (a)} - \text{Prob. of error in (b)}}{\text{Prob. of error in (a)}}. \tag{7}$$

The second statistic involves a similar calculation for predicting C given E. In addition to this asymmetrical statistic, Goodman and Kruskal (1963) also proposed a symmetrical one [28] (see Appendix A).

In psychology, meteorology, and clinical medicine, the *kappa statistic* [29] has often been used as a measure of agreement, which is a special case of association (Cohen, 1960; Heidke, 1926, described in Brier & Allen, 1951). Let π_0 indicate the observed probability of agreement, and π_E indicate the probability of chance coincidence when the two variables are independent (see Appendix A). The kappa statistic is then defined as follows:

$$\kappa \triangleq \frac{\pi_0 - \pi_E}{1 - \pi_E}. \tag{8}$$

This index has been used as a measure of "response agreement" in psychology (Cohen, 1960, 1968; Fleiss, 1971; Light, 1971), "observer agreement" in clinical medicine (Landis & Koch, 1977), and the "accuracy" of weather forecasting (called the *Heidke skill score*) especially in America (Woodcock, 1976) because it can be regarded as the proportion correct (*PC*) [14] corrected for chance success. Woodcock also introduced another measure of accuracy in weather forecasting, the *skill test* [30], S_k, which is also a corrected version of the proportion correct.

Some researchers (e.g., A. W. F. Edwards, 1963; Goodman, 1970) have proposed the *log odds ratio* of diagonal cell frequencies as a measure of association in a 2×2 contingency table. However, this index coincides with Yule's (1900) Q provided it is standardized using the method shown in Appendix A.

4.4. Normative models in the philosophy of science

Mill's (1843/1973) *method of difference* is based on the idea that a cause is the difference between cases where the effect did occur and cases where the effect did not occur. Cheng and

Novick (1992) proposed a probabilistic interpretation of this idea, but as shown in Appendix A, this is formally equal to "converse" ΔP (i.e., ΔP^c [3]).

In his attempt to construct a theory of temporal direction, Reichenbach (1956) defined an asymmetric measure of causality. According to his definition, "an event B is *causally between* C and E" (p. 190) if (a) $0 < P(E) < P(E|C) < P(E|B) < 1$, (b) $0 < P(C) < P(C|E) < P(C|B) < 1$, and (c) $P(E|C,B) = P(E|B)$. He did not characterize this relation quantitatively. When considering just the cause and effect, this measure is equivalent to judging the existence of a causal relation by comparing $P(E|C)$ with $P(E)$, and $P(C|E)$ with $P(C)$. From Equation 3, it is clear that this idea is very close to the phi-coefficient, the normative index of correlation in statistics.

In an attempt to formalize causality in networks of events, Good (1961, 1962) defined a measure $Q(E{:}C)$, which is the "causal support for E provided by C" or the "tendency of C to cause E" [31,32]. The index was defined as the log odds ratio of the conditional probabilities when the effect is absent: $P(\bar{E}|\bar{C})$ and $P(\bar{E}|C)$. Again this index needs to be standardized as it runs over the interval $[-\infty, \infty]$ (see Appendix A).

Suppes (1970) called an event, C, a *prima facie cause* when (a) the effect, E, occurred after C; (b) $P(C) \neq 0$; and (c) $P(E|C) > P(E)$. Note that the condition c is implied by Reichenbach's (1956) condition a. Although Suppes also did not introduce any quantitative measures, Cheng and Novick (1992) defined an index based on Suppes' condition c as follows [33,34]:

$$S \stackrel{\Delta}{=} P(E|C) - P(E). \tag{9}$$

4.5. Descriptive models in psychology

Klayman and Ha (1987) argued that *positive hypothesis testing* (comparing cell a with b) and *positive target testing* (comparing cell a with c) are favored in contingency judgment when the probability of the effect is small, $P(E) < .5$, and the probabilities of the cause and effect are similar, $P(C) \approx P(E)$. Here, we define an index based on this idea that runs over the interval [0, 1], which we called the *positive test* index [15]:

$$PT \stackrel{\Delta}{=} \frac{P(E|C) + P(C|E)}{2}. \tag{10}$$

McKenzie (1994) identified some further heuristic indices in the literature on human covariation assessment. Some of them are among the indices that have already been introduced, and others are regarded as specific forms of a more general parameterized model, the so-called *weighted linear combination model* (Anderson & Sheu, 1995; Einhorn & Hogarth, 1986; Schustack & Sternberg, 1981):[4]

$$\beta_0 + \beta_1 a + \beta_2 b + \beta_3 c + \beta_4 d. \tag{11}$$

First, the *positive hits* strategy [16], L_1, is realized by proposing that $\beta_1 = 1$ and all others are 0. Second, the *hits minus false positives* strategy [17,18], L_2, is realized by proposing that $\beta_1 = 1$, $\beta_2 = -1$, and all others are 0. Third, the *sum of diagonals* strategy [19], L_3, is realized by proposing that $\beta_1 = \beta_4 = 1$, $\beta_2 = \beta_3 = -1$, and $\beta_0 = 0$. This index concerns the diagonal

cell frequencies and is similar to Inhelder and Piaget's (1958) association coefficient and the log odds ratio index introduced earlier. It is sometimes called ΔD and has been used many researchers as an index of causal strength or contingency (Allan & Jenkins, 1980, 1983; Arkes & Harkness, 1983; Jenkins & Ward, 1965; Kao & Wasserman, 1993; Shaklee & Hall, 1983; Shaklee & Mims, 1982; Shaklee & Tucker, 1980; Ward & Jenkins, 1965; Wasserman et al., 1990). Finally, McKenzie proposed a new index called the *aggregate model* [20, 21], L_4, to model differential cell impact. This is another version of the linear combination model, with $\beta_1 = 4$, $\beta_2 = -3$, $\beta_3 = -2$, $\beta_4 = 1$, and $\beta_0 = 0$. All versions of the linear combination model are sensitive to sample size.

We now compare indices for 2×2 contingency tables including the dual factor heuristic. However, before exhaustively compare the indices using past experimental data, an experiment was conducted to establish the criteria for judging whether to include or exclude data from each study in the literature.

5. Experiment 1: Discrete versus continuous task

As experiments in the literature are varied in methods with different motivations, appropriate standards are important for a meta-analysis. Lax criteria cause heterogeneity in data, which can generally lead to vague or even false conclusions. Our main concern is the discrimination between the *discrete* and the *continuous* paradigm (Anderson & Sheu, 1995) in causal judgment experiments. In the discrete paradigm experiments, participants only observe a sequence of either presence or absence of the cause and effect in pairs with particular frequencies set by the experimenter in advance as a form of 2×2 contingency table. In the continuous paradigm, which is sometimes called a *free-operant probabilistic schedule*, people choose between *doing* and *not doing* the cause at every moment and the effect occurs according to the probabilities previously set by the experimenter.

As we pointed out when we introduce the dual factor heuristic, the difference between these experiment paradigms is directly related to the recently much emphasized distinction between *observation* and *intervention* in human causal inference (Lagnado & Sloman, 2004; Steyvers et al., 2003). The discrete paradigm, being based solely on observation, is likely to only engage the heuristic stage of causal judgment. The continuous paradigm involves active intervention on behalf of participants and is hence likely to mainly enage the analytic stage. As these paradigms engage different processing stages they should produce different behavior and if this is the case, we should not mix up data from the two types of tasks in the subsequent meta-analyses.

In this experiment, we used H and the phi-coefficient (ϕ) to examine the difference in performance between the two paradigms. As H is regarded as a simpler substitute (omitting d-cell) of the well-known normative index of correlation, ϕ, comparing the data fit between H and ϕ would be a reasonable test to assess the difference in people's sensitivity to d-cell frequencies in the discrete and continuous paradigms.[5] Because in the continuous paradigm the data sequence is partly under participant control, it is difficult to fully equate experiments using these different paradigms. However, given a range of possible contingencies, we would expect H to do better than ϕ for the discrete task where only observation is allowed, whereas

for the continuous task, where intervention focuses attention of causal necessity, we would expect ϕ to do at least as well if not better than H.

5.1. Method

5.1.1. Tasks

Participants were asked to assess the strength of the causal relation between using a particular type of fertilizer and plants blooming. In the *Discrete Task*, participants only observed a sequence of scenes (a total of 12–15) in which fertilizer (cause) and plant blooming (effect) were either present or absent. The cell frequencies of the 2×2 contingency tables used in this task are shown in Table 3 (left). Each participant was presented with 12 stimuli in a randomized order, each corresponding to a line in the table.

In the *Continuous Task*, two probabilities were assigned per stimulus, as shown in Table 3 (right): a probability that the plant blooms (E) when fertilizer was used (C), $P(E|C)$; and a probability that the plant blooms (E) when fertilizer was not used ($\bar{C}$), $P(E|\bar{C})$. Each participant inspected 30 cases for each stimulus. In each case, he or she chose to either use or not use fertilizer and observed plants bloom or not bloom. Given participant's choices, the plant bloomed according to the probabilities in Table 3 (right).

5.1.2. Procedure

The experiment was conducted on personal computers. Every time a participant clicked the mouse, a new picture was displayed in a randomized order. After observing a series of situations (i.e., pictures), participants rated the subjective strength of the causal relation with a value between 0 (completely unrelated) and 100 (completely related). This cycle was repeated for all stimuli shown in Table 3.

Table 3
Stimuli in discrete (Left) and continuous (Right) tasks

No	a	b	c	d	H	ϕ	No	$P(E\|C)$	$P(E\|\bar{C})$
1	2	8	1	4	.37	.00	1	.20	.20
2	1	4	2	8	.26	.00	2	.50	.20
3	5	5	1	4	.65	.29	3	.50	.50
4	2	2	2	8	.50	.30	4	.80	.20
5	4	4	2	2	.58	.00	5	.80	.50
6	2	2	4	4	.41	.00	6	.80	.80
7	8	2	1	4	.84	.58			
8	4	1	2	8	.73	.58			
9	8	2	2	2	.80	.30			
10	4	1	5	5	.60	.29			
11	8	2	4	1	.73	.00			
12	4	1	8	2	.52	.00			

5.1.3. Participants and design

A total of 39 undergraduate students from Ritsumeikan University participated in this experiment as unpaid volunteers. They were randomly assigned either to the Discrete Task or to the Continuous Task. The participants were run either individually or in small groups.

5.2. Results and discussion

Figure 1 shows the relation between index values and participant ratings. In the Continuous Task, contrary to the Discrete Task, each participant experienced different stimuli (i.e., cell configurations). This is because the experimenter cannot control how many times a participant

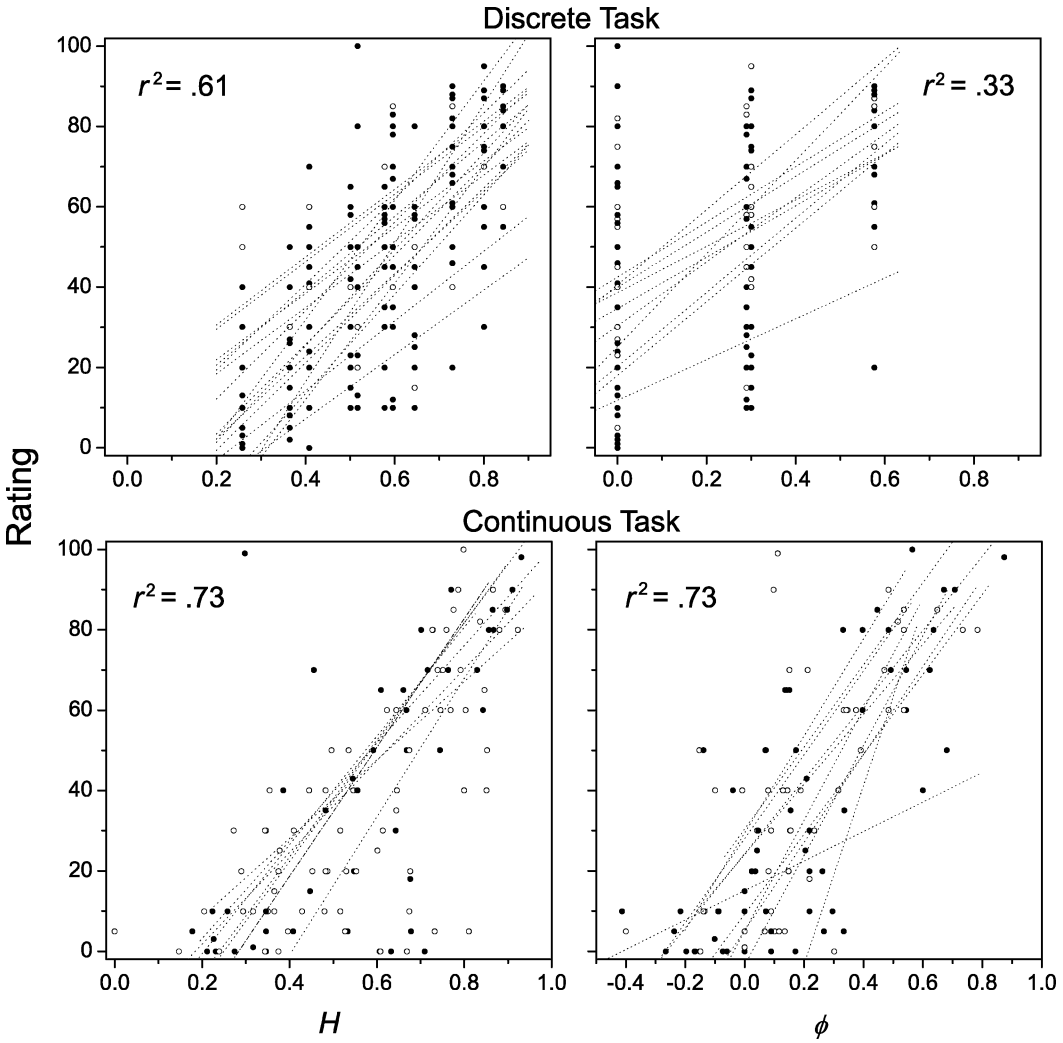

Fig. 1. Participants' ratings of causal strength in Discrete (upper panels) and Continuous (lower panels) Tasks of Experiment 1. Regression lines are for participants whose ratings significantly correlated with the index ($p < .05$).

performs the action that is the possible cause of the effect and partly because the experimenter can only control the probability with which the effect occurs consequent on the action and the actual occurence is left to chance. Consequently, every rating is plotted as a single point in Fig. 1. In this figure, data for participants whose correlation coefficients were significant ($p < .05$) are plotted with black circles and others are plotted with white ones. Regression lines were drawn for individual participants (only for those participants whose ratings had a significant correlation with the index). Correlation coefficients were calculated by participants and averaged using Fisher's z transformation.

In the Discrete Task, the correlation of participants' ratings was much stronger with H ($r^2 = .66$) than with ϕ ($r^2 = .33$). Thus, when only observational data is available people tend to ignore d-cell information. In the Continuous Task, participant ratings had as much correlation with ϕ ($r^2 = .73$) as with H ($r^2 = .73$). Therefore, with the continuous paradigm where intervention focuses attention on causal necessity, as predicted, the correlation with ϕ rose and was similar to H, indicating as much analytic as heuristic processing in this paradigm.

The results showed differences in causal cognition between situations where people only observe the occurrences of events and situations where they intervene in the system. Consequently, continuous paradigm data were excluded in the following meta-analyses assessing the merits of the dual factor heuristic as an account of the heuristic stage.

However, the dual factor heuristic, which is a model of the non-interventional heuristic stage in causal induction, well described the data on the Continuous Task, as did ϕ. This result suggests that continuous tasks are not completely analytic. Some participants may have processed the data only at the heuristic level. Our explanation of some d-cell effects in covariation detection suggests that there may be individual differences in strategy usage. Moreover, a reviewer suggested that to control for inevitable differences between the continuous and discrete paradigms, we could have used a yoking procedure. In this procedure, the sequence of trials generated by a participant in the continuous paradigm is summarized and presented to a yoked participant in the discrete paradigm. These are clearly both areas for future research which we hope to turn to in the future.

6. Meta-analysis 1: Non-parameterized models

Many experiments have investigated human and animal judgments on causality, correlation, or contingency between two covarying events. However, as far as we know, no study has exhaustively examined the descriptive validity of the various models based on covariation listed in Table 2. The purpose of this section is to begin to determine the most appropriate descriptive model for covariation detection based on 2×2 contingency tables.

6.1. Method

Experiments on judgments of causality, correlation, or contingency that satisfy the following five criteria were selected from the literature: (a) The participants were human; (b) the causal relation in the experiment was assumed to be a one-to-one correspondence (i.e., a single event causes an effect); (c) the task was to assess the subjective strength of a relation with a

(nearly) continuous value; (d) the candidate cause and effect were presented sequentially with frequencies decided in advance by the experimenter (i.e., the discrete paradigm experiments); (e) the task handled ordinary generative causes (as opposed to preventive ones that suppress the effect). Although the criterion d was justified by Experiment 1, criterion e requires further justification.

Some experiments (e.g., Buehner, Cheng, & Clifford, 2003) impose a generative or preventive scenario as different tasks, whereas some experiments (e.g., Wasserman et al., 1996) merge generative and preventive causes within a task. A problem is that the recontextualization required to interpret preventive causes (discussed in the section, *The Dual Factor Heuristic*) can differ between these experiments, and we have no way of determining any recontextualization participants may impose to interpret preventive causes.

Moreover, learning preventive associations from purely observational data without recontextualization would be computationally intractable and is another instantiation of the frame problem we discussed in introducing the dual factor heuristic. As we move around the world most things are not happening. Do we therefore develop associations between the occurrence of an event (a possible cause) and all the other events that are not occurring at the time, just in case they may be instances of a preventive cause? This would involve encoding an enormous number of largely spurious associatons every time an event was encountered between a representation of that event and a representation of almost all other possible events that are not occurring (i.e., the negation of each other possible such event). In sum, at the level of learning covariations to feed into subsequent interventional processing in the real world, as opposed to the laboratory, it would be computationally infeasible for people to attend to preventive associations. Consequently, to avoid data heterogeneity we only included experiments that measured positive relations. We acknowledge that these exclusions may give the impression that we are only testing H against a very specialized set of data. However, while specialized with respect to the classes of experimental paradigms used in the area, we would argue that it is *not* specialized with respect to people's real world experience of detecting covariations by observing the world.

The following six target experiments were identified in the literature: Experiment 1 of Anderson and Sheu (1995), abbreviated to "AS95"; Experiments 1 and 3 of Buehner, et al. (2003),[6] abbreviated to "BCC03.1" and "BCC03.3," respectively; Experiments 1 through 3 of Lober and Shanks (2000),[7] abbreviated to "LS00"; and Experiments 2 and 6 of White (2003), abbreviated to "W03.2" and "W03.6," respectively. In addition, Experiment 1 of the current study was also included in the analysis. For each experiment, all indices mentioned in the previous section and the converse indices (indicated by the superscript c in Table 2) were calculated from the cell frequencies. In total, 34 indices were calculated. To measure each model's fit to the data, the coefficient of determination (r^2) between each index and participants' mean ratings of causal strength were calculated.

6.2. Results and discussion

The results of the analysis are set out in Table 4. As an overall measure of goodness-of-fit, the weighted average of the seven values of r^2 for each index was calculated and they are shown in column 11 (i.e., "whole [w/o Expt 2]"). This calculation was based on the Fisher z

Table 4
Data fit (r^2) for the non-parameterized models

	Data									
Model	AS95	BCC03.1	BCC03.3	LS00	W03.2	W03.6	Expt 1	Expt 2	whole (w/o Expt 2)	whole
1 H	.91**	.94**	.91**	.80**	.69*	.80	.93**	.96**	.90	.91
2 ΔP	.78**	.86**	.70*	.77**	.00	—	.50*	.00	.75	.72
3 ΔP^c	.78**	.74**	.68*	.56**	.36	—	.50*	.73**	.73	.73
4 PW	.30**	.75**	.90**	.41**	.15	—	.55**	.01	.39	.36
5 PW^c	.53**	.43*	.48	.25	.01	—	.49*	.95**	.47	.52
6 PN	.56**	.35*	.37	.27	.09	—	.37*	.00	.47	.44
7 PN^c	.41**	.68**	.83*	.38*	.09	—	.38*	.13	.43	.39
8 SP	.81**	.88**	.74*	.23**	.18	.45	.79**	.36	.77	.76
9 SDT	.78**	.48**	.68*	.23	.00	—	.47*	.00	.68	.65
10 SDT^c	.78**	.48**	.68*	.23	.00	—	.47*	.00	.68	.65
11 R	.76**	.86**	.70*	.77**	.49	.18	.16	.49*	.74	.72
12 C	.82**	.63**	.82*	.17	.20	.54	.00	.80**	.76	.73
13 C^c	.58**	.30	.23	.19	.01	.28	.34*	.97**	.48	.53
14 PC	.76**	.86**	.70*	.77**	.49	.18	.49*	.16	.74	.72
15 PT	.89**	.90**	.88**	.75**	.51*	.72	.90**	.80**	.87	.87
16 L_1	.46**	.63**	.82*	.13	.57*	.86	.73**	.85**	.50	.53
17 L_2	.73**	.63**	.82*	.15	.66*	.72	.85**	.89**	.71	.72
18 L_2^c	.49**	.86**	.70*	.75**	.14	.40	.42*	.84**	.55	.57
19 L_3	.71**	.86**	.70*	.75**	.49	.18	.49*	.01	.71	.67
20 L_4	.86**	.93**	.99**	.78**	.95**	.88	.91**	.43	.88	.87
21 L_4^c	.78**	.96**	.90**	.88**	.59**	.70	.76**	.60*	.81	.81
22 ϕ	.78**	.84**	.70*	.71**	.21	—	.50*	.49*	.75	.74
23 χ^2	.07*	.83**	.63	.69**	.26	—	.48*	.36	.23	.24
24 Q	.78**	.49**	.67*	.23	.00	—	.47*	.01	.68	.64
25 Y	.79**	.50**	.59	.18	.00	—	.49*	.07	.68	.64
26 $\lambda_{E\|C}$	.22**	.61**	.29	.57**	—	—	.36*	.45*	.30	.30
27 $\lambda_{C\|E}$	.08*	.86**	.70*	.77**	.37	—	.36*	.56*	.26	.27
28 λ	.13**	.82**	.65	.71**	.37	—	.36*	.49*	.29	.30
29 K	.71**	.86**	.70*	.77**	.39	—	.51**	.65**	.71	.71
30 S_k	.72**	.86**	.70*	.77**	.28	—	.50*	.73**	.71	.71
31 G	.70**	.74**	.89**	.38*	.13	—	.58**	.04	.66	.63
32 G^c	.54**	.92**	.89**	.77**	.78**	—	.59**	.88**	.65	.66
33 S	.49**	.86**	.70*	.77**	.00	—	.31	.11	.53	.48
34 S^c	.61**	.30	.23	.19	.00	—	.31	.44	.49	.49

Note: AS95: Experiment 1 of Anderson and Sheu (1995), the number of stimulus variation, $M = 80$; BCC03.1 and BCC03.3: Experiments 1 and 3 of Buehner et al. (2003), $M = 13$ and 6; LS00: Experiments 1–3 of Lober and Shanks (2000), $M = 11$; W03.2 and W03.6: Experiments 2 and 6 of White (2003b), $M = 8$ and 4; Expt 1 and Expt 2: Experiments 1 and 2 of the current study, $M = 12$ and 9.
In the "whole" column, the figures indicate the weighted average of r^2 (see the text for details).
** $p < .01$, * $p < .05$.

transformation of r and each study was weighted by its degree of freedom, $M - 3$, where M indicates the number of stimuli used in the experiment (e.g., see, Rosenthal, 1991, chap. 4). The number of stimuli, M, was 80, 13, 6, 11, 8, and 4 for AS95, BCC03.1, BCC03.3, L00, W03.2, and W03.6, respectively.

As shown in Table 4, only four indices exceeded .80 in average r^2 value; these were H (.90), L_4 (.88), PT (.87), and L_4^c (.82). Five indices including the top four and G^c ($r^2 = .65$) were statistically significant ($p < .05$) in all the data (with the exception of W03.6, which had a low N), although $G^c$was ranked low (21/34). Below the top four, some indices including ϕ and ΔP are grouped in the range of .77 to .65. We also assessed the model fits using non-parametric, rank order correlations (i.e., Spearman's coefficient of correlation; e.g., see Siegel & Castellan, 1988). Our findings were essentially unaltered. The same indices (i.e., H,L_4, PT, and L_4^c) remained as the distinctive top four.

H and PT are of course conceptually very similar: H is the geometric mean of $P(E|C)$ and $P(C|E)$ and PT is the arithmetic mean of these two factors. H has the advantage of being a limiting case of the normative index (i.e., it has a rational motivation). PT is obviously going to be highly correlated with H and we can only understand its success to the extent that it approximates H.

L_4 and its converse L_4^c are based on setting the parameters of the weighted linear combination model to particular values. However, unlike the other indices produced in this way, the parameter values have been specifically set, albeit only crudely, to capture the differential cell weightings found in previous data. That L_4 and L_4^c provide good fits to these new data sets without allowing these parameters to be further adjusted is impressive. It demonstrates the generalizability of the model. However, these indices are not based on a normative theory (i.e., they provide no explanation for why people behave as they do in these tasks). The only explanation they provide is the tautological one that they can fit the previous data. In contrast, H is based on a limiting case of the normative index of covariation in 2×2 contingency tables when C and E are rare. Thus, H *explains* why people behave as they do (i.e., it is an adaptively rational strategy; Chater, Oaksford, Nakisa, & Redington, 2003). Thus, we argue that H is to be preferred because it not only provides a good fit to the data, it also rationally explains why people behave in the way they do on these tasks.

7. Experiment 2: H versus ΔP and L_4

Although the results of the meta-analysis indicated that the dual factor heuristic model best fit the experimental data, other models did rather well. However, these degrees of fit may have occurred because some indices are correlated with H in the stimulus sets used in the experiments against which we evaluated the models in the above meta-analysis. More specifically, when H fits the data, and H and a particular index are dependent for some stimuli, the index can also show a good fit to the data. As long as stimuli are used for which different models' predictions do not diverge, we cannot differentiate between models. The motivation for Experiment 2 was therefore to present stimulus set where these indices are predicted to diverge in their predictions.

We focused our attention on ΔP and L_4. First, because L_4 and its converse fit the data well, we needed to differentiate these indices from H. Second, because ΔP has been one of the leading psychological models of causal induction and it did quite well in the last meta-analysis ($r^2 = .75$, ranked 8th), ΔP was worth comparing with H.[8] So far, almost all experiments on

causation and correlation based on covariation information have been concerned with ΔP in one way or another, either positively or negatively (e.g., Allan, 1980, 1993; Allan & Jenkins, 1980, 1983; Anderson & Sheu, 1995; Baker, Berbrier, & Vallée-Tourangeau, 1989; Buehner et al., 2003; Chapman & Robbins, 1990; Cheng, 1997; Cheng & Novick, 1992; Griffiths & Tenenbaum, 2005; Neunaber & Wasserman, 1986; Shaklee & Tucker, 1980; Shanks, 1987; Vallée-Tourangeau, Murphy, Drew, & Baker, 1998; Wasserman, 1990; Wasserman, Chatlosh, & Neunaber, 1983; Wasserman et al., 1993; White, 2000, 2003). Third, as mentioned in the previous section, although *PT* matched H, *PT* is less rationally motivated and it showed a (slightly) worse fit. Consequently, we do not consider *PT* further.

This experiment was therefore designed to examine which model best predicts the data in a stimulus set where H is not well correlated with these other indices. The cell frequencies were explicitly defined such that H and ΔP had independent values on three levels (low, middle, and high) and no data could validate both models at the same time. This procedure also had the consequence that neither L_4 nor its converse was significantly correlated with H.

7.1. Method

7.1.1. Participants

Fifty undergraduate students of Ritsumeikan University participated in this experiment as unpaid volunteers. The participants were run either individually or in small groups of up to 10.

7.1.2. Stimuli and procedure

Table 5 indicates the cell frequencies and the values of each index for the nine contingency tables used in this experiment. The experiment was a completely within-participant design conducted on personal computers. Each participant was presented with stimuli derived from these nine contingency tables in randomized order.

Table 5
Stimuli in Experiment 2

Cell frequency					Models' prediction			
a	b	c	d	N	H	ΔP	L_4	L_4^c
1	1	12	28	42	.20	.20	5	−6
1	0	24	24	49	.20	.50	−20	−44
1	0	19	78	98	.22	.80	44	25
3	1	6	5	15	.50	.20	2	−3
3	1	6	18	28	.50	.50	15	10
1	0	3	12	16	.50	.80	10	7
12	3	3	2	20	.80	.20	35	35
12	3	3	7	25	.80	.50	40	40
9	1	3	27	40	.82	.80	54	52

Note: $N = a + b + c + d$.

The cause was "drinking milk" and the effect was "stomach-ache." Each contingency table in Table 5 corresponded to a person, who could have a weak digestion. Participants were instructed to judge the causal strength between drinking milk and stomach-ache for each contingency table. When a participant clicked the mouse, a pair of pictures was displayed: The picture on the left of the screen indicated whether the person drank the milk, and the picture on the right of the screen indicated whether the person experienced a stomach-ache. After observing a series of these stimuli (i.e., paired pictures) in randomized order, participants rated the subjective strength of the causal relation with a value from 0 (*not related at all*) to 100 (*completely related*) using the computer mouse. This procedure was repeated for the nine contingency tables.

The exact instructions were as follows:[9]

> The purpose of this task is to investigate how strongly people feel a cause produces its effect.
>
> Imaginary persons you will see on the screen may have a stomach-ache. You are asked to judge whether milk is the "cause" of their stomach-ache after observing several situations in which they drink or do not drink milk. There is no correct answer, so please answer according to your purely subjective feeling.
>
> This task consists of several sessions and each session is concerned with a different person. When the session starts, every time you click the mouse, a new picture will be displayed indicating (1) whether or not the person drank milk, and (2) whether or not the person had a stomach-ache. After watching some cases, you must decide the degree to which you think milk is the "cause" of a stomach-ache for that person. Please use the computer mouse to indicate your response. Rate as a rough estimate from 0 (there is no causal relation at all between milk and a stomach-ache) to 100 (there is a complete causal relation). You will be asked to do this several times.

7.2. Results and discussion

Figure 2 shows participants' mean ratings of causal strength. In Figure 2a, the x axis corresponds to the value of H for each contingency, and each line connects the points that have the same ΔP value. It is clear that the changes in participants' causal evaluations bear no relation to the ΔP values. On the other hand, it is clear that there is a linear relation between H and causal strength. An analysis of variance (ANOVA) with the causal ratings as the dependent variable and the three levels of H as a factor showed a highly significant main effect of the dual factor heuristic, $F(2, 98) = 274.30$, $p < .0001$, but no main effect of ΔP, $F(2, 98) = 0.42$, $p = .67$. The interaction between these two factors was also significant, $F(4, 98) = 5.62$, $p < .001$. As Figure 2a shows when $H = .5$ and $H = .8$, there are differences in causal ratings between levels of ΔP. However, these were inconsistent. When $H = .8$, causal ratings increased with ΔP, as would be expected; whereas, when $H = .5$, causal rating decreased with ΔP, which is the exact opposite of what would be expected.

In Figure 2b, the x axis corresponds to the value of L_4 or L_4^c, and points are connected with lines in groups of three that had the same value of H. Figure 2b shows that there is no clear monotonic relation between participants' ratings and L_4 or L_4^c. We can see that L_4 and L_4^c make poor predictions particularly for the group of stimuli when $H = .2$ where there is almost no correlation between causal ratings and L_4 (or L_4^c). A similar ANOVA could not be

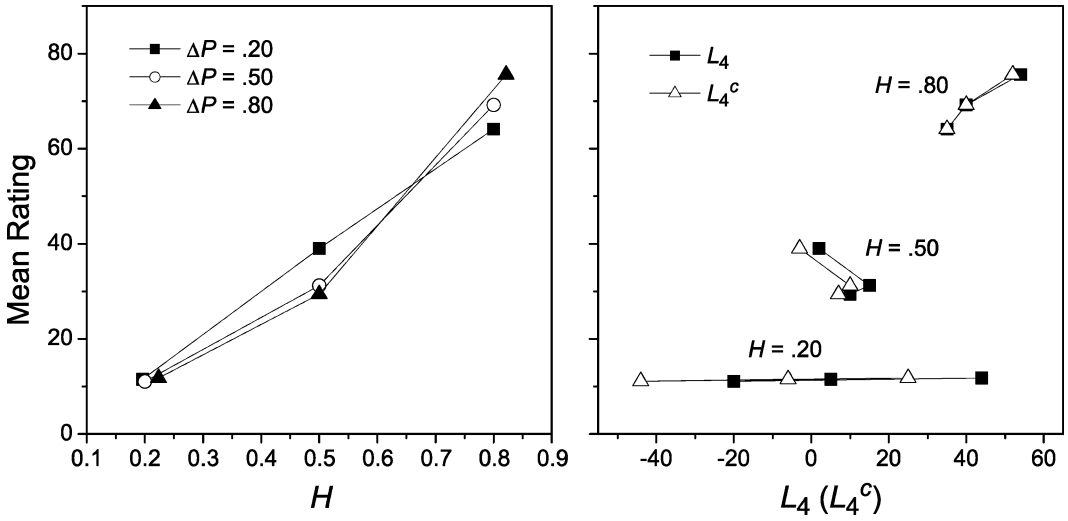

Fig. 2. Participants' ratings of causal strength and the predictions of (a) H and ΔP and (b) the predictions of L_4 and L_4^c in Experiment 2.

carried out for these indices, because the levels were not fully crossed. However, Figure 2b replicates the odd interaction effects we observed for ΔP (i.e., when $H = .8$, causal ratings increase; but when $H = .5$, causal ratings decrease as L_4 or L_4^c increase).

The "Expt 2" column of Table 4 shows r^2 for this experiment. H had an r^2 of .96, ΔP had an r^2 of .00, L_4 had an r^2 of .43, and its converse had an r^2 of .60. An analysis using Spearman's non-parametric correlation coefficient replicated the results: .93, .00, .40, and .72 for H, ΔP, L_4, and L_4^c, respectively. The rightmost column of Table 4 shows the weighted averages of r^2 based on Fisher's z transformation (as described in the section, *Meta-analysis 1*) for all the data including Experiment 1 and 2. Here we can see that H still provides the best fit. These results suggest that in the meta-analysis H actually fit the data, whereas ΔP and L_4 seemed to fit the data because of the correlation in these stimuli between H and these other two indices.

We also wondered how H compared to the fully parameterized weighted linear combination model when an index of fit that penalizes model complexity (i.e., the number of free parameters) is used. We therefore performed a second meta-analysis, which also included a range of other parameterized models of covariation detection that have been prominent in the literature and that, therefore, needed to be compared to H.

8. Meta-analysis 2: Parameterized models

We have now established that H compares well with other non-parameterized models of covariation detection. It outperforms other models including normative and descriptive models of covariation detection. Moreover, in an experiment where the other best performing models were not well correlated with H, H performed a lot better than those models. However, we

also needed to compare H with the range of parameterized models of covariation detection in the literature. This is the purpose of the second meta-analysis that we now report.

8.1. Parameterized models

As we indicated above, the first parameterized model we consider is the weighted linear combination model, L_* [41] (Equation 11) introduced in the section *Indices for 2 × 2* Contingency Tables, which is one of the most familiar models with free parameters. However, a range of other models exist in the literature.

8.1.1. The simple Bayesian model

Another covariational approach was proposed by Anderson and Sheu (1995; Anderson, 1990) as a rational analysis of causal judgments. Causal strength was proposed to be a function of the odds of a causal relationship (i.e., a hypothesis, H) given the data (D), $O(H|D)$:

$$B \stackrel{\Delta}{=} \frac{O(H|D)}{1 + O(H|D)}. \tag{12}$$

Here, $O(H|D)$ is described using Bayesian theorem, letting p_c be the probability of the effect in the presence of a cause, p_a be the probability of the effect in the absence of the cause, p_n be a base rate of the effect irrespective of occurrence of the cause, and p_r be the prior probability of the causal relationship, $P(H)$;

$$\begin{aligned} O(H|D) &= \frac{P(H|D)}{P(\bar{H}|D)} = \frac{P(H)}{P(\bar{H})} \cdot \frac{P(D|H)}{P(D|\bar{H})} \\ &= \frac{p_r\, {p_c}^a (1-p_c)^b\, {p_e}^c (1-p_e)^d}{(1-p_r)\, {p_n}^{a+c} (1-p_n)^{b+d}}. \end{aligned} \tag{13}$$

They called this parameterized model the *Simple Bayesian Model* [40].

8.1.2. Weighted ΔP models

Several models with free parameters have also been proposed in the associative learning literature. According to this approach, causal judgments are explained by learning theory where the candidate cause is treated as the CS, the effect is treated as the US, and causal strength is treated as associative strength. The Rescorla–Wagner model (Rescorla & Wagner, 1972) is expressed as a simultaneous difference equation so that it can predict the results of step-by-step learning. However, the Rescorla–Wagner model involving one causal factor (contrasted with the context) reduces to the following formula at asymptote (Wasserman et al., 1993), which is sometimes called weighted ΔP, ΔP_{w1} [36]:

$$\Delta P_{w1} \stackrel{\Delta}{=} \frac{w_a a}{w_a a + w_b b} - \frac{w_c c}{w_c c + w_d d} = \frac{a}{a + w_1 b} - \frac{c}{c + w_2 d}. \tag{14}$$

The right hand side of this equation is derived by reductions assuming both w_a and w_c are not 0. Using this measure to capture the fit between the data and the Rescorla–Wagner model assumes that performance in each experiment is at asymptote. For the studies we model, this criterion cannot be guaranteed to hold. However, no simple criterion like number of

trials could guarantee the opposite (i.e., that performance is *not* at asymptote). Consequently, although admitting that this *may* be a limitation, we continued to model all the data that met our selection criteria (we discuss whether performance is at asymptote further in the section *General Discussion*.) Confusingly, there is another version of weighted ΔP (Anderson & Sheu, 1995), ΔP_{w2} [37], which has been applied to causal judgment in humans:

$$\Delta P_{w2} \stackrel{\Delta}{=} \beta_0 + \beta_1 P(E|C) - \beta_2 P(E|\bar{C}). \tag{15}$$

8.1.3. Information integration models

Kao and Wasserman (1993) derived the same equation as ΔP_{w1} from Busemeyer's (1991) information integration theory. They also proposed that a non-normative information integration process can be described as follows [38]:

$$\begin{aligned} I_1 &\stackrel{\Delta}{=} w_0 + \frac{w_a a - w_b b - w_c c + w_d d}{w_a a + w_b b + w_c c + w_d d} \\ &= w_0 + \frac{a - w_1 b - w_2 c + w_3 d}{a + w_1 b + w_2 c + w_3 d}, \end{aligned} \tag{16}$$

which is a form obtained by substitutions, $w_1 = w_b \,/\, w_a$, $w_2 = w_c \,/\, w_a$, and $w_3 = w_d \,/\, w_a$. Also based on Busemeyer's theory, Anderson and Sheu (1995) introduced another model [39]:

$$\begin{aligned} I_2 &\stackrel{\Delta}{=} .5 + w_1 \frac{a}{N} - w_2 \frac{b}{N} - w_3 \frac{c}{N} + w_4 \frac{d}{N} \\ &= \beta_0 + \beta_1 \frac{a}{N} - \beta_2 \frac{b}{N} - \beta_3 \frac{c}{N}. \end{aligned} \tag{17}$$

This is derived by substituting $(.5 + w_4)$, $(w_1 - w_4)$, $(w_2 + w_4)$, and $(w_3 + w_4)$ with β_0, β_1, β_2, and β_3, respectively.

8.1.4. The Pearce model

Pearce (1987) proposed another associative model of stimulus generalization. Asymptotic predictions for his model are made according the following formula (Buehner et al., 2003; Perales & Shanks, 2003), J [35]:

$$J \stackrel{\Delta}{=} \frac{x_2[a(c+d) - c(ax_1 + bx_1)]}{a(c+d) + b(c+d) - cx_1(ax_1 + bx_1) - dx_1(ax_1 + bx_1)}. \tag{18}$$

8.1.5. Parameterized dual factor heuristic

The dual factor heuristic, H, has no parameters, but to compare it with these other parameterized models, we used the simplest linear equation as a response function which maps strength of covariation onto an actual response: $\beta_0 + \beta_1 H$.

8.1.6. Causal support with the power transformation

Griffiths and Tenenbaum (2005) proposed a transformation of the causal support model [8], *SP*, based on the power law to accommodate non-linearity of the data before fitting their model as follows: $\beta_0 + \beta_1 \text{ sign}(SP) \text{ abs}(SP)^{\gamma}$.

8.2. Method

The aforementioned models were compared using the same set of data as Meta-analysis 1 including data from Experiment 1 plus additional data from Experiment 2. The model parameters were estimated based on the least squared error criterion. For models that have the form of linear equations (i.e., ΔP_{w2} [37], I_2 [39], L_* [41], and parameterized H [1]), a multiple linear regression was conducted.[10] For the other non-linear models (i.e., J [35], ΔP_{w1} [36], I_1 [38], B [40], and transformed SP [8]), *nonlinear least squares regression* (Dennis, Gay, & Welsch, 1981) was conducted using the "nlregb" function of S-PLUS 6.0.

In all experiments, the causal judgments were rated between 0 and 100, whereas the output of all the aforementioned indices is between 0 and 1. Therefore, all the definitions for the models were multiplied by 100. (This only affects the absolute value of AIC_c [see below], not the relative magnitude which is essential.)

8.3. Results and discussion

As a measure of goodness-of-fit, r^2 between participants' causal ratings and the best fit predictions based on the least squared error criterion was calculated for each model and they are shown in Table 6. Estimated parameters are shown in Appendix B. In the "whole" column of Table 6, the weighted averages of r^2 (see the section *Meta-analysis 1*) are shown. J, ΔP_{w1}, and ΔP_{w2} fit the data from W03.2 and Experiment 2 poorly ($r^2 \leq .20$), but the other models seemed to fit all the data quite well.

However, with parameterized models, the r^2 goodness-of-fit index is generally not an adequate criterion. The more free parameters a model possesses the more chance there is that it will fit the data. However, an "overcomplex" model loses *generalizability,* which is the ability to fit samples beyond the current data set (e.g., see Pitt & Myung, 2002). As a measure of generalizability, for each model we therefore calculated AIC_c (Hurvich & Tsai, 1989), which is a corrected version of the well-known Akaike Information Criterion (AIC; Akaike, 1973). AIC_c is generally appropriate and dramatically outperforms AIC in small-sample settings such as those in our meta-analyses (e.g., see Burnham & Anderson, 1998). Smaller values of AIC_c indicate that a model is more generalizable. According to this measure, H was always ranked in the top three for all data sets and overall it is more generalizeable than the other parameterized models.[11] The last column of Table 6 shows the mean rank orders of AICc. AICc values for all indices could not be derived for BCC03.3 and W03.6 as they contained insufficient stimuli. These experiments were therefore excluded from the averaging process. Again, H was the best among these eight parameterized models according to mean rank order of AICc, which was significant by Friedman's test, $\chi_r^2(7) = 19.0$, $p = .015$.

In sum, even compared with complex parameterized models, the dual factor heuristic performed well. Note that although the number of parameters was taken into account in this comparison, the number of cells (or the number of samples) used for the calculation to detect covariation was not. All indices other than H use information about all cells, while H ignores d-cell information. Generally speaking, the less information needed for a calculation, the easier it is to perform, which means a lower cognitive load is imposed. This is an intrinsic advantage of the dual factor heuristic not factored in to the model comparison process.

Table 6
Data Fit (r^2) and generalizability (AIC_c) for the parameterized models

		AS95		BCC03.1		BCC03.3		LS00		W03.2		W03.6	Expt 1		Expt 2		whole (mean)	
Model	K	r^2	AIC_c	r^2	AIC_c	r^2	AIC_c	r^2	AIC_c	r^2	AIC_c	r^2	r^2	AIC_c	r^2	AIC_c	$\bar{r}^2$	rank
1 H	(2)	.91**	259.3†	.94**	46.8††	.91**	43.6††	.80**	57.7†	.69**	40.5†††	.80	.93**	47.3†††	.96**	38.2†††	.91	1.8
35 J	2	.89**	302.6	.92**	63.2	.87**	49.0†	.87**	52.7††	.14	52.6	.54	.89**	54.4††	.00	69.0†	.85	4.5
36 ΔP_{w1}	2	.93**	309.9	.92**	64.4	.93**	42.7†††	.91**	49.1†††	.07	52.0	.54	.89**	54.5†	.03	70.3	.89	5.0
37 ΔP_{w2}	3	.94**	222.3†††	.96**	46.4†††	.99**	59.5	.87**	57.9	.20	57.3	.54	.89**	57.6	.00	75.4	.92	4.3
8 SP	(3)	.80**	321.7	.95**	48.5†	.76*	79.5	.71**	66.6	.66*	50.5	.45	.77**	66.5	.44*	70.1	.80	6.3
38 I_1	4	.72**	351.3	.94**	57.6	.99**	—	.88**	64.0	.99**	43.6††	—	.92**	59.9	.72**	76.1	.82	6.3
39 I_2	4	.93**	242.6††	.96**	51.9	.99**	—	.87**	65.2	.97**	48.7	—	.92**	59.4	.85**	70.5	.93	4.8
40 B	4	.88**	282.6	.96**	53.3	.96**	—	.89**	63.2	.97**	48.5†	—	.94**	56.6	.91**	66.1††	.91	3.8
41 L_*	5	.86**	297.4	.96**	59.4	.99	—	.89**	74.6	.97**	104.7	—	.93**	67.4	.98**	77.3	.91	8.0

Note: K indicates the number of parameters in each model. With regard to the number of parameters of the model H, see the text. For W03.6, because AIC_c could not be calculated for all indices because of the lack of stimulus variation (i.e., 4), this column was omitted.

In the "whole" column, r^2 was averaged with each study weighted as in Meta-analysis 1 and mean rank orders of AIC_c are shown. BCC03.3 and W03.6 were excluded from the averaging because of the insufficiency of stimulus variation.

$^{**}p < .01$, $^{*}p < .05$; ††† best model, †† second best model, † third best model.

To investigate this feature in more detail, further computer simulations were conducted to provide a fuller rational analysis of covariation detection using the dual factor heuristic.

9. Simulation 1: Effectiveness and parsimony

The results of the meta-analyses and the experiment show that the dual factor heuristic has some descriptive validity. It would appear that people use this heuristic as an approximation to the normative strategy of covariation assessment when they assess causal strength between a candidate cause and its effect.

In the simulations we now report, we attempted to provide a more detailed rational analysis of the dual factor heuristic. According to Anderson (1990), the adaptive rationality of a cognitive process depends on the limitations imposed by working memory and the nature of the environment. We considered three factors that are relevant to the adaptive rationality of the dual factor heuristic: rarity, equiprobability, and the capacity of working memory. The importance of rarity has already been established in deriving the dual factor heuristic. The rarity assumption was initially introduced in explaining data selection behavior in the Wason selection task (Oaksford & Chater, 1994). In the same task, Hattori (2002) also proposed that participants make a *biconditionality assumption* that a conditional is often regarded as a biconditional (i.e., not only is "if X then Y" true so is "if Y then X"). To avoid misunderstanding, we call this assumption the *equiprobability assumption* here because it suggests that the probabilities of the antecedent, X, and the consequent, Y, of an indicative conditional are almost equal (see also, Klayman & Ha, 1987). Working memory limitations have been appealed to in many rational analyses (e.g., Anderson, 1990). Interestingly, for our current analyses, Kareev (2000) has pointed out that the working memory limitations are not always disadvantageous to cognitive processing. He pointed out that restricting the sample size within the narrow limits imposed by working memory (i.e., approx. 7 ± 2) may amplify a sample correlation so enabling people to detect correlations more efficiently.

We first explored how well and under what conditions H best predicts the normative index of covariation. This first analysis shows that H is most *effective* as an approximation to the phi-coefficient when only small samples are used (i.e., it is most effective when parsimony is enforced by working memory limitations).

9.1. Method

In these simulations we generated a large variety of contingency tables embodying different relations between cause and effect. To generate these tables we needed to vary three parameters that specify a population of tables. We chose $P(C)$, $P(E)$ and $P(C,E)$. The probability of the candidate cause, $P(C)$, and the effect, $P(E)$, were both varied systematically over the range .1 to .9 in steps of .1 to examine the effects of rarity and equiprobability. Their joint probability was set by a random variable with a unified distribution defined between $P(C)P(E)$ and $P(C)$ or $P(E)$ (depending which is the smaller): $P(C,E) \sim Unif(P(C)P(E), \min(P(C),P(E)))$. These are the parameters of the population of contingency tables.

The number of samples was controlled in two ways: One was the total sample size (i.e., $N = a + b + c + d$); the other was to exclude the d-cell (i.e., $N_W = a + b + c$); and the sample size was varied systematically to examine the effects of working memory capacity, which is known to be about 7 ± 2: N (or N_W) $\sim Norm(\mu, (\mu/7)^2)$, $\mu = (7, 14, 28, 56)$. To be specific, when the number of samples is controlled by N_W, the procedure was as follows. First, $P(C)$, $P(E)$, and μ were set at particular levels systematically. For each set of $P(C)$, $P(E)$, and μ, the correlation between H and ϕ was calculated based on the generated contingency tables. For each contingency table, $P(C,E)$ and N_W were determined according to their respective probability distributions. Each sample's cell category (a, b, c, or d) was determined sequentially according to the following probabilities (as the parameters of the population), $P(C,E)$, $P(C) - P(C,E)$, $P(E) - P(C,E)$, $1 - P(C) - P(E) + P(C,E)$, respectively. Sampling was continued until the number of samples categorized in a, b, or c (excluding d) summed up to N_W. As a result, the real sample size exceeds N_W by exactly the size of the d-cell.

For each set of $P(C)$, $P(E)$, and μ, 500 contingency tables were generated, regenerating $P(C, E)$ and N (or N_W) for each table. For each contingency table, H and ϕ were calculated, and finally r^2 between H and ϕ was calculated. To obtain stable estimates, this procedure was repeated 20 times and each r^2 was averaged.

9.2. Results and discussion

9.2.1. Relations between H and sample ϕ

Figure 3a shows r^2 between H and ϕ when the equiprobability assumption is made. The horizontal axis indicates the mean probability of cause or effect and the coordinates are linked by sample size. The results have two distinctive features. First, when the probability of events is low (i.e., rare), the correlation between H and sample ϕ is always high, irrespective of

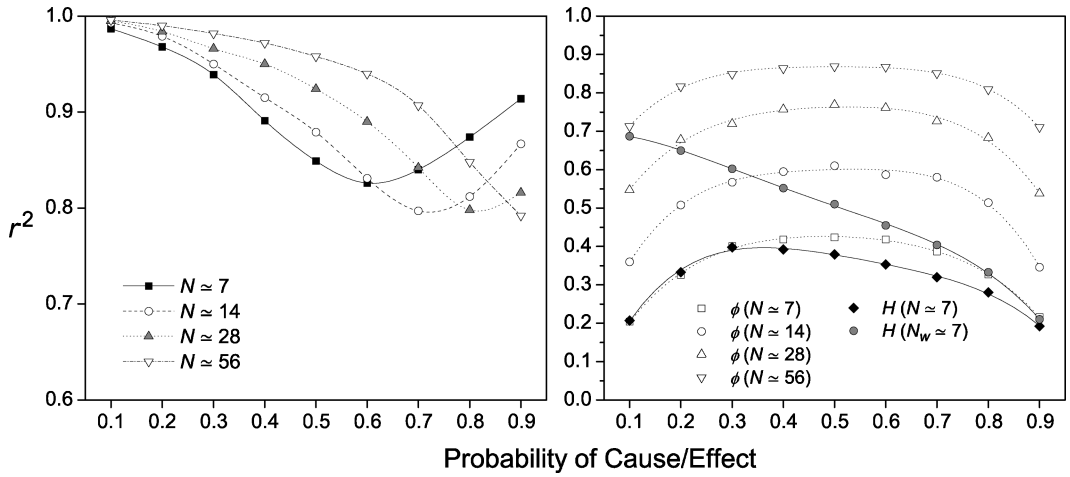

Fig. 3. (a) Coefficients of determination (r^2) between H and ϕ when equiprobability is maintained; (b) coefficients of determination between H and ϕ and between ϕ and ϕ when equiprobability is maintained.

sample size. Second, as the probability of events increases, the correlation between H and sample ϕ decreases.

H differs from normative indices in not using the frequency information of the d-cell. The results, however, show that non-normative H works well and it works as well as ϕ given some constraints. Moreover, H may be preferable because it minimizes the computational cost of calculating an estimate of causal strength (i.e., H is more economical than ϕ).[12]

9.2.2. Relations between H *and population* ϕ_0

Sampling does not aim at comprehension of the sample itself. For an agent living in the real world, it is essential to estimate parameters of the population from samples. Consequently, population ϕ is significantly more important than sample ϕ. From now on, to avoid possible confusion, population ϕ is described as ϕ_0, and when ϕ appears alone, it indicates sample ϕ.

When estimating the population ϕ_0, sample ϕ provides the best estimates of ϕ_0 under the multiple sampling model (e.g., Bishop et al., 1975), and the accuracy of the estimates increases as the number of samples increases. Figure 3b shows the precision with which H and ϕ predict population ϕ_0 for various sample sizes. In this figure, ϕ is plotted on dotted lines and H is plotted with black diamonds on a line in the small sample case ($N \approx 7$). These results show that estimation accuracy of ϕ_0 by H matches that for ϕ when $P(C)$ is in the domain .1 to .3 (i.e., rare).

However, because H ignores the d-cell frequency, more samples can be acquired than other models that use all four cells—that is, the capacity limitation of working memory forces the heuristic to control the number of samples such that N_W (not N) is approximately equal to 7. The results of sampling by this method are shown in Figure 3b with a line and grey circles, which shows that H predicts ϕ_0 far more precisely than ϕ in the domain where the probability of cause or effect is .1 to .5. In particular, when the probability is low (approximately .1–.3), N_W-based H achieves higher accuracy than ϕ with double the sample size (i.e., approximately. 14).

Figure 4 shows that H predicts ϕ_0 more accurately when equiprobability is maintained. The horizontal axis indicates the difference of probabilities of the cause and the effect, $P(C) - P(E)$. Equiprobability occurs at the point where this value is equal to 0. Points that have the same mean probability (m) of the cause and the effect are connected by a line. For instance, if $P(C)$ and $P(E)$ are (.1, .3), (.2, .2), or (.3, .1), the mean probability of cause and effect is .2, so they are connected with a line "$m = .2$" and the values of "$P(C) - P(E)$" of these points are –.2, .0, and .2, respectively. This figure shows the results when $N_W \approx 7$ and $P(C)$, $P(E) \leq .5$ (i.e., when rarity is maintained). Figure 4 shows clearly that H predicts ϕ_0 far better when both events are equiprobable.

10. Simulation 2: Efficiency

In the real world, people usually sample sequentially. In sequential sampling, it is important to form an appropriate conclusion quickly. Accordingly, there is a trade-off between the early cessation of sampling and the accuracy of estimation. Therefore, we investigated the relation between the convergence of an index value on the true population mean and sample size.

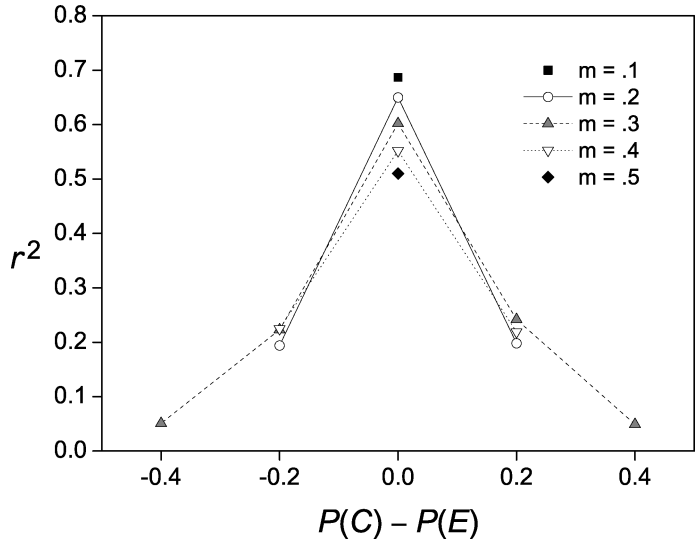

Fig. 4. Coefficients of determination between H and ϕ_0 under the rarity assumption and with $N_W = a + b + c \approx 7$. M indicates the mean probability of the cause (C) and the effect (E), and horizontal lines indicate the degree of deviation from equiprobability.

10.1. Method

The convergence of ϕ, ΔP, and H was examined in sequential sampling from a population. The probabilistic characteristics of the population were defined by three parameters, $P(C)$, $P(E)$, and ϕ_0.[13] Assuming rarity and equiprobability, $P(C)$ and $P(E)$ were set to .2, and ϕ_0 was varied at three levels, .2, .5, and .8. The sample size started at 1 and increased to 30. This procedure was repeated 5,000 times, and the mean values of the indices at each N (or N_W) and their standard deviations were derived.

10.2. Results and discussion

The results are shown in Figure 5 for the cases where $\phi_0 = .8$ (upper) and $\phi_0 = .2$ (lower). The results for $\phi_0 = .5$ were similar to these two and so were omitted. The y-axis in each upper panel indicates the mean values of ϕ, ΔP, and H for 5,000 trials of sequential sampling. These show that the indices converge as the number of samples increases. The lower panels for each level of ϕ_0 show the *SD*s for each index. The dotted line with black diamonds indicates N_W-based index values, whereas the other lines are N based. When ϕ_0 was .2 and .5, the results were the same.

The mean values of H and ϕ converge gradually. As the limiting value of H is higher than that of ϕ, the convergence line for H is always above that of ϕ; the rates of convergence themselves, however, do not differ. On the other hand, ΔP appears to be stable from the start, as does N_W-based H.

The reliability of an index cannot be determined solely from the convergence rate of the mean. It is necessary to know the confidence interval of the population mean. As the standard

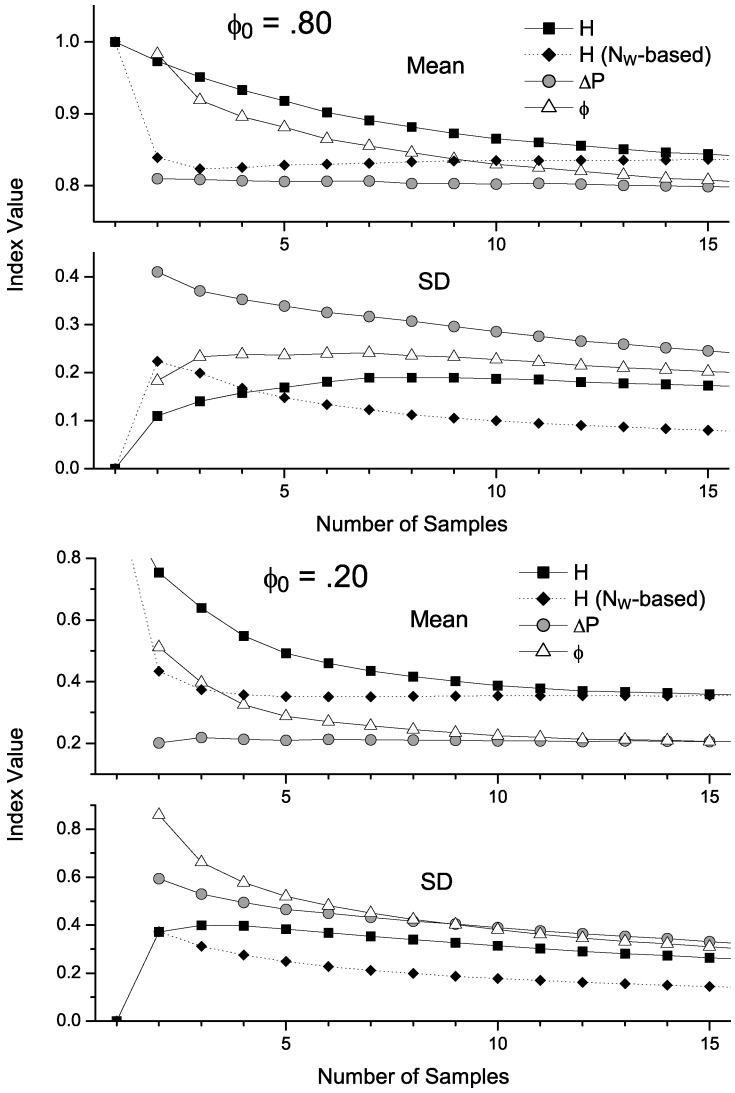

Fig. 5. Convergence graphs for the sample indices (ϕ, ΔP, and H) as functions of sample size when $\phi_0 = .8$ (upper) and $\phi_0 = .2$ (lower) with 5,000 trials of sequential sampling. The horizontal axes indicate the number of samples, N or N_W (detailed in the text); the vertical axes indicate the mean or the standard deviation of each index (i.e., ϕ, ΔP, or H).

deviation decreases, the confidence interval narrows and the information about the population gained from an index increases. The convergence status of the standard deviations indicates that ΔP, which seemed very stable when considering its mean convergence alone, has large standard deviations and turns out not to be a reliable index. On the other hand, N_W-based H has small standard deviations. Taken together with its rate of mean convergence this clearly indicates that N_W-based H is a more reliable index, assuming rarity and equiprobability.

10.3. Summary of simulations

In an environment where rarity and equiprobability hold, the index H defined by the dual factor heuristic measures ϕ approximately. In such an environment, H can function as a superior substitute for ϕ. When H is used for causal induction, d-cell information can be disregarded; more samples can thus be acquired within the limits of working memory. Using this N_W-based sampling, H can predict the population ϕ_0 far better than other indices, even better than sample ϕ (Simulation 1). In sequential sampling, the confidence interval of H decreases rapidly, and it is the most reliable predictor of population ϕ_0 (Simulation 2).

11. General discussion

The dual factor heuristic, H, provided the best fit to the experimental data both from meta-analyses of past data and new experiments. Normative assessments of the strength of covariation require attending to all events, including those that involve nothing happening (d-cell). Attending to all cell frequencies places prohibitive demands on people's limited working memory capacity. Computing H allows people to assess the strength of relations without using the d-cell. This strategy is more efficient on memory and is accurate under circumstances where $P(E)$ and $P(C)$ are small. Despite the model's descriptive power and adaptive rationality, d-cell neglect is controversial. In this discussion we first consider the d-cell effects.

Given an observation of small but reliable positive d-cell effects (e.g., Anderson & Sheu, 1995), the dual factor heuristic might be viewed as falsified at birth. However, the conclusion of this research is that the dual factor heuristic best approximates people's behavior in the heuristic stage of causal induction. The results do not imply either that there are no d-cell effects at all or that any attempts to investigate the effects are in vain. There are many other factors besides cell frequencies that can influence covariation assessment—for example, context, intelligence, depression, culture, gender, and so on. Despite the success of the dual factor heuristic in its restricted domain of application, we do not pretend that these factors have no effect on causal judgment. The same is true of the d-cell effects. The d-cell effects may be a real part of causal induction despite the descriptive validity of the dual factor heuristic. If the effects are significant, it might suggest either that people adopt a slightly different algorithm from this model, that these effects arise at the analytic stage, or that some assumptions of the model are slightly wrong.

Nonetheless, our simulations showed that H can perform even better than using the d-cell if one is only allowed to keep track of a small number of events in working memory. Our simulations also revealed some further factors that influence the model with respect to adaptation to the real world. Next, we consider the results of the simulations indicating the importance of rarity, equiprobability, and working memory capacity to covariation detection and cognition more generally.

There is ample empirical evidence for working memory limitations but rather fewer demonstrations of the adaptive advantage such an apparent limitation can confer on a cognitive agent. Consequently, our demonstration that covariation estimation using H can be more accurate

than using normative measures for small samples is consistent with related demonstrations by Kareev (2000). The idea that causal statements usually assume low probability events that are approximately equal has also occurred in other domains such as hypothesis testing (Klayman & Ha, 1987). Moreover, there is recent evidence that when people spontaneously frame conditional statements about causal hypotheses (i.e., *if cause then effect*), they do so using rare events (see McKenzie et al., 2001).

Rarity and equiprobability have also emerged as important factors in other areas, in particular, in Oaksford and Chater's (1994) information gain model of data selection (see also Hattori, 2002). Suppose that there is a certain category Y in question, and an arbitrary category X. Let $\mathbf{X} = \{X, \bar{X}\}$ and $\mathbf{Y} = \{Y, \bar{Y}\}$, and let $I(\mathbf{Y})$ describe average *self-information* (i.e., *entropy*) of a random variable $\mathbf{Y}$. The reduced entropy (i.e., information gain) of $\mathbf{Y}$ by knowing $\mathbf{X}$ is

$$I(\mathbf{Y}) - I(\mathbf{Y}|\mathbf{X}), \tag{19}$$

which has local maxima when (a) $P(X) = P(Y) = P(X, Y)$, and (b) $P(X) = 1 - P(Y)$ and $P(X, Y) = 0$. The former corresponds to equiprobability between X and Y, and the latter corresponds to equiprobability between $\bar{X}$ and Y. If the rarity assumption is made, case a is the only solution. Therefore, when an explanatory category X is searched for the category Y, seeking a category that has approximately the same probability increases the expected information gain. Thus the assumptions that allow H to provide a good index of causal strength are same as those made by *optimal data selection models* in explaining the rationality of people's data selection behavior (Hattori, 2002; Oaksford & Chater, 1994, 2003).

This study contrasts with most other rational analyses of causal induction (e.g., Anderson, 1990; Griffiths & Tenenbaum, 2005) insofar as it is concerned with an optimal *approximation* in an uncertain world. Other rational analyses (e.g., Anderson, 1990; Oaksford & Chater, 1994) showed that people's behavior is adaptive if the environment can be assumed to have certain characteristics (e.g., rarity or a power law need probability function). However, the analyses themselves are able to make predictions for behavior even if these environmental assumptions are violated. Oaksford and Chater (1994), speculating on the consequences for the algorithmic level, suggested two extremes. On the one hand, the cognitive system could implement the rational analysis in a hard wired heuristic such that, for example, the system shows no behavioral variation with respect to violations of rarity. On the other hand, the analysis could be directly implemented in the cognitive system in which case behavior should perfectly track rarity violations. Oaksford and Chater (1994) suggested that the truth probably lies somewhere between these two extremes: Although there may be some responsiveness to rarity violation, the cognitive system would be expected to show some inertia in responding to deviations away from normal environmental assumptions. In the dual factor heuristic, the environmental assumption about rarity is built in (i.e., it cannot respond to rarity violation and thus can only approximate rational performance). Therefore, it is closer to the hard-wired heuristic end of Oaksford and Chater's (1994) continuum.

This fact may question the rationality of using this heuristic. However, if the assumptions of the model are usually respected in the environment then in the large majority of cases it will provide the right answers quickly and efficiently. The question that obviously arises is, can the cognitive system adapt to cases of rarity violation? That there is some d-cell sensitivity and McKenzie and Mikkelsen's (2007) recent work suggests that people may have some

facility to adapt to violations of rarity. Besides, although the dual factor heuristic provides a rational strategy, this does not necessarily mean that all people detect covariation in exactly the same way as the model, but it does mean people's average behavior can be predictable. Each individual may make judgments in a different way. Not only individual differences but also context, wording, or general knowledge may affect causal induction (e.g., Arkes & Harkness, 1983; Crocker, 1982; Einhorn & Hogarth, 1986).

There is an intrinsic difference between an associational view of covariation judgment and our heuristic view. The dual factor heuristic is designed to make quick responses to the environment when there is pressure to continuously construct provisional judgments. On this view, the learning process is too slow to suggest tentative hypotheses for test by intervention at the analytic level. For example, when someone feels sick, they want to detect causally relevant food as soon as possible to avoid possible future risks. According to the dual factor heuristic, a single case in sequential sampling can have a strong impact if it is an instance of the a-cell. On the other hand, according to the associative approaches, the impact of instances is small and the effect on judgment is gradual and incremental. It is highly unlikely that people always reserve their judgments until learning reaches "asymptote" as the associationists insist. However, sample size no doubt increases the reliability of data, which can also be important in some context, perhaps where Type I errors (i.e., false alarms) matter decisively. For example, if you were a manager of a baseball team you would be unlikely to hire a player based on observing him hit a single, albeit decisive, home run. Rather you would assess his overall performance based on his batting average because the costs of a false positive are too high (given the high earnings of these players). Assessing the costs of false positives and so whether long run performance must be assessed is again a matter for the analytic stage of causal induction not the first pass determination of likely causal candidates.

Moreover, there is a phenomenon that has been regarded as a "bias" from an associative view and that seems to have a rational explanation in our model. In *outcome density bias* the probability of the outcome, $P(E)$, for non-contingent relations (i.e., $\Delta P = 0$) can positively affect human judgments of causal strength (e.g., Allan & Jenkins, 1980, 1983; Chatlosh, Neunaber, & Wasserman, 1985; Shanks, 1985; Wasserman & Shaklee, 1984). This bias has often been observed in the continuous paradigm tasks. Although the continuous paradigm is not in the scope of the current study, this bias may be explained, at least partly, by the dual factor heuristic. H can be written as follows:

$$H = \frac{P(C)[1 - P(C)]}{\sqrt{P(C)P(E)}}\Delta P + \sqrt{P(C)P(E)}. \tag{20}$$

When $\Delta P = 0$, H is equal to $\sqrt{P(C)P(E)}$. Suppose that participants "perform" the act (i.e., the candidate cause) that may cause the effect irrespective of the outcome density. We can thus regard $\sqrt{P(C)}$ as a fixed constant, k, and so, $H = k\sqrt{P(E)}$. This means that outcome density, $P(E)$, should alter people's estimates of the predictive relationship (i.e., the higher the outcome density the stronger the perceived causal relation).

In conclusion, when people observationally detect covariation between events as a first step to induce causality, their behavior is most consistent with the dual factor heuristic which is a simple and efficient strategy that approximates the normative index, assuming rarity. At present, participants' performance in a particular type of causal induction task is best

described by this heuristic, which suggests that people's behavior in non-interventional causal assessment tasks is adaptively rational.

Notes

1. The method McKenzie (1994) adopted was not the Monte Carlo method because it was not based on random variables with some probabilistic distributions.
2. Although ϕ distinguishes positive and negative correlations in any 2×2 contingency tables by means of the "$\pm$" sign of the index value, H does not have such a mechanism—that is because, obviously, $\lim_{a,d\to 0} \phi = -1$, whereas $\lim_{a,d\to 0} H = 0$.
3. In terms of his theory, it is characterized by *monotonicity* and *exogeneity*. It is called monotonic if and only if there is no individual case that already has the effect without the cause but that would lose the effect if it gained the cause. It is called exogenous if and only if the cause and the effect are not influenced by any common factors.
4. Schustack and Sternberg (1981) actually included an additional term for the strength of competing causes and Equation 11 is according to Anderson and Sheu (1995).
5. Note that fitting a linear regression model of cell frequencies and looking for different cell weights is not an appropriate test. See the third point in the section *What d-cell Insensitivity is Not*, in this regard.
6. In Experiment 1 of Buehner, Cheng, and Clifford (2003), Stimuli 11 (8, 0, 8, and 0 as a, b, c, and d, respectively) and 15 (0, 8, 0, and 8, ditto) were omitted because many indices (11 and 13 indices out of 40 for Stimuli 11 and 15, respectively) cannot be calculated according to the zero divisor. As a result, the number of stimuli counted for this experiment amounted to 13.
7. In this study, the variation of stimuli in each experiment was very small (i.e., 3 or 4), and their configurations were set up to complement each other, so we figure these three experiments as one experiment here.
8. Although ΔP seems to be invalidated by the data of W03.2 (i.e., $r^2 = .0$), this study only set two levels for ΔP, and so it might be seen insufficient evidence to negate this index.
9. These instructions were written in Japanese.
10. For some model-data combinations (i.e., L_* with BCC03.1, BCC03.3, LS00, and W03.02; and I_2 with BCC03.1, BCC03.3, and LS00), we had to compute rank-deficient linear least squares solutions. In such cases, the Moore–Penrose generalized inverse of a matrix was used (see Venables & Ripley, 1999, p. 100).
11. Because $\mathrm{AIC_c}$ is already adjusted not only for the number of parameters, K, included in a model but also for the sample size (i.e., the number of stimuli, M, included in each experiment), unlike the case of r^2, it was not weighted by the number of stimuli as an overall measure of the model's appropriateness.
12. It is controversial how to stipulate computational simplicity. For more argument on this point, see for example, Chater, Oaksford, Nakisa, and Redington (2003).

13. Between $P(C, E)$ and ϕ_0, there is a relation as $P(C, E) = xy + \phi_0\sqrt{x\bar{x}y\bar{y}}$, where $x = P(C)$, $y = P(E)$, $\bar{x} = 1 - x$, and $\bar{y} = 1 - y$.

Acknowledgments

A part of this study was conducted while the authors were at the School of Psychology, Cardiff University, Wales. We appreciate the school's warm support for our research. Preparation of this article was partially supported by Grant-in-Aid for Scientific Research 19500229 from the Japan Society for the Promotion of Science, a research grant from Institute of Human Sciences, Ritsumeikan University (Project Research B), and a grant from The Daiwa Anglo-Japanese Foundation awarded to M. Hattori. This article was presented, in part, at the 4th International Conference on Cognitive Science (ICCS 2003), The University of New South Wales, Sydney, Australia.

We thank Marc Buehner, Kyung Soo Do, Ken Manktelow, Masanori Nakagawa, Minoru Nakashima, Tatsuo Otsu, Tsuneo Shimazaki, Tetsuo Takigawa, Hiroshi Yama, and Kimihiko Yamagishi, as well as Josh Tenenbaum and three anonymous reviewers for their very helpful and constructive comments on this study. We are also grateful to Shiori Nakao, Yuka Otake, Tomoko Tamezane, and Miyuki Tanaka for their help in running the experiments.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the 2nd international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

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

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

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

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

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

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

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

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

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

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America, 53*, 129–160.

Brier, G. W., & Allen, R. A. (1951). Verification of weather forecasts. In T. F. Malone (Ed.), *Compendium of meteorology: Prepared under the direction of the committee on the compendium of meteorology* (pp. 841–848). Boston: American Meteorological Society.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1119–1140.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory: Volume 1 cognition* (pp. 187–215). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes, 90*, 63–66.

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

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

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

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

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Crocker, J. (1982). Biased questions in judgment of covariations studies. *Personality and Social Psychology Bulletin, 8*, 214–220.

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

Dennis, J. E., Jr., Gay, D. M., & Welsch, R. E. (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software, 7*, 348–368.

Edwards, A. W. F. (1963). The measure of association in a 2×2 table. *Journal of the Royal Statistical Society: Series A (General), 126*, 109–114.

Edwards, J. H. (1957). A note on the practical interpretation of 2×2 tables. *British Journal of Preventive and Social Medicine, 11*, 73–78.

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

Evans, J. St.B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Evans, J. St.B. T., & Over, D. E. (1996a). *Rationality and reasoning*. Hove, England: Psychology Press.

Evans, J. St.B. T., & Over, D. E. (1996b). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review, 103*, 356–363.

Finley, J. P. (1884). Tornado predictions. *American Meteorological Journal, 1*, 85–88.

Fisk, A. D., & Schneider, W. (1984). Memory as a function of attention level of processing and automatization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 181–197.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378–382.

Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretationi of confirmation bias phenomena. *Psychological Review, 100*, 298–319.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford, England: Oxford University Press.

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming diculties in bayesian reasoning: A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychological Review, 106*, 425–430.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1–67.

Good, I. J. (1961). A causal calculus (I). *The British Journal for the Philosophy of Science, 11*, 305–318.

Good, I. J. (1962). A causal calculus (II). *The British Journal for the Philosophy of Science, 12*, 43–51.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classfications. *Journal of the American Statistical Association, 65*, 226–256.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–764.

Goodman, L. A., & Kruskal, W. H. (1963). Measures of association for cross classifications, III: Approximate sampling theory. *Journal of the American Statistical Association, 58*, 310–364.

Green, D. M., & Swets, J. A. (1966/1988). *Signal detection theory and psychophysics*. Los Altos, CA: Peninsula. (Original work published 1966)

Griffths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384.

Hanssen, A. W., & Kuipers, W. J. A. (1965). On the relationship between frequency of rain and various meteorological parameters. *Mededelingen en Verhandelingen, 18*, 2–15.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*, 208–216.

Hattori, M. (2001). Ingakihou-no niyouin heuristic model [A dual-factor heuristic model of causal induction]. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, 8*, 444–453.

Hattori, M. (2002). A quantitative model of optimal data selection in Wason's selection task. *The Quarterly Journal of Experimental Psychology, 55A*, 1241–1272.

Heidke, P. (1926). Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst [Calculation of the success and goodness of strong wind forecasts in the storm warning service]. *Geografiska Annaler, 8*, 301–349.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Hurvich, C., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures* (A. Parsons & S. Milgram, Trans.). London: Routledge & Kegan Paul.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79*, 1–17.

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

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review, 107*, 397–402.

Klauer, K. C. (1999). On the normative justification for information gain in Wason's selection task. *Psychological Review, 106*, 215–222.

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

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 856–876.

Landis, J. R., & Koch, G. G. (1975a). A review of statistical methods in the analysis of data arising from observer reliability stuies (Part 1). *Statistica Neerlandica, 29*, 101–123.

Landis, J. R., & Koch, G. G. (1975b). A review of statistical methods in the analysis of data arising from observer reliability stuies (Part 2). *Statistica Neerlandica, 29*, 151–161.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76*, 365–377.

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

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York: Wiley.

Mandel, D. R., & Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General, 127*, 269–285.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In D. Michie (Ed.), *Machine intelligence* (Vol. 4, pp. 463–502). Edinburgh: Edinburgh University Press.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategires: Covariation assessment and Bayesian inference. *Cognitive Psychology, 26*, 209–239.

McKenzie, C. R. M., Ferreira, V. S., Mikkelsen, L. A., McDermott, K. J., & Skrable, R. P. (2001). Do conditional hypotheses target rare events? *Organizational Behavior and Human Decision Processes, 85*, 291–309.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin and Review, 7*, 360–366.

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology, 54*, 33–61.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*, 440–466.

Mill, J. S. (1843/1973). *A system of logic ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation* (Vols. 7, 8). Toronto: University of Toronto Press. (Original work published 1843).

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.

Neunaber, D. J., & Wasserman, E. A. (1986). The effects of unidirectional versus bidirectional rating procedures on college students' judgments of response–outcome contingency. *Learning and Motivation, 17*, 162–179.

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning, 2*, 1–31.

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

Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind & Language, 6*, 1–38.

Oaksford, M., & Chater, N. (1993). Reasoning theories and bounded rationality. In K. I. Manktelow & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 31–60). London: Routledge.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Oaksford, M., & Chater, N. (1995). Theories of reasoning and the computational explanation of everyday inference. *Thinking and Reasoning, 1*, 121–152.

Oaksford, M., & Chater, N. (Eds.). (1998a). *Rational models of cognition*. Oxford, England: Oxford University Press.

Oaksford, M., & Chater, N. (1998b). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Hove, England: Psychology Press.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin and Review, 10*, 289–318.

Over, D. E., & Green, D. W. (2001). Contingency, causation, and adaptive inference. *Psychological Review, 108*, 682–684.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *Adaptive decision maker*. Cambridge, England: Cambridge University Press.

Pearce, J. M. (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. (2003). Normative and descriptive accounts of the influence of power and contingency on causal judgement. *The Quarterly Journal of Experimental Psychology, 56A*, 977–1007.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*, 421–425.

Pylyshyn, Z. W. (Ed.). (1987). *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Reichenbach, H. (1956). *The direction of time*. Berkeley: University of California Press.

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

Rescorla, R. A., & 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 research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Thousand Oaks, CA: Sage.

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

Shaklee, H., & Hall, L. (1983). Methods of assessing strategies for judging covariation between events. *Journal of Educational Psychology, 75*, 583–594.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 208–224.

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

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., & Dickinson, A. (1987). Associative accounts of causality judgment. *The Psychology of Learning and Motivation, 21*, 229–261.

Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 265–311). London: Academic.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.

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

Smedslund, J. (1963). The concept of correlation in adults. *The Scandinavian Journal of Psychology, 4*, 165–173.

Sosa, E., & Tooley, M. (Eds.). (1993). *Causation*. Oxford, England: Oxford University Press.

Sperber, D., Premack, D., & Premack, A. J. (Eds.). (1995). *Causal cognition: A multidisciplinary debate*. Oxford, England: Oxford University Press.

Stanovich, K. E. (1999). *Who is rational? Studies of individual dierences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99*, 100–117.

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401–409.

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

Toda, M. (1983). Future time perspective and human cognition: An evolutional view. *International Journal of Psychology, 18*, 351–365.

Vallée-Tourangeau, F., Murphy, R. A., Drew, S., & Baker, A. G. (1998). Judging the importance of constant and variable candidate causes: A test of the power PC theory. *The Quarterly Journal of Experimental Psychology, 51A*, 65–84.

Venables, W. N., & Ripley, B. D. (1999). *Modern applied statistics with S-PLUS* (3rd ed.). New York: Springer-Verlag.

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

Wasserman, E. A. (1990). Detecting response–outcome relations: Toward an understanding of the causal texture of the environment. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 27–82). New York: Academic.

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

Wasserman, E. A., Dorner, W. W., & Kao, S.-F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 509–521.

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.

Wasserman, E. A., Kao, S.-F., Van Hamme, L. J., Katagiri, M., & Young, M. E. (1996). Causation and association. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 207–264). London: Academic.

Wasserman, E. A., & Shaklee, H. (1984). Judging response–outcome relations: The role of response–outcome contingency, outcome probabiolity, and method of information presentation. *Memory & Cognition, 12*, 270–286.

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

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

Woodcock, F. (1976). The evaluation of yes/no forecasts for scientific and administrative purposes. *Monthly Weather Review, 104*, 1209–1214.

Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt.

Yule, G. U. (1900). On the association of attributes in statistics: With illustrations from the material of the childhood society. *Philosophical Transactions of the Royal Society of London: Series A, 194*, 257–319.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75*, 579–642.

Appendix A: Mathematical supplements on the indices

A.1. Causal support

Griffiths and Tenenbaum (2005) defined *causal support* by the value of the log likelihood ratio for obtaining data D from G_C over G_I, each of which is a causal graphical model (see Equation 6). $P(D|G_C)$ is defined by integrating over all possible values p_C and p_B could assume. Assuming independence between C and B and a uniform prior distribution for p_B, it is simplified as shown in the following equation using the beta function and cell frequencies

(a, b, c, d, and $N = a + b + c + d$):

$$P(D|G_I) \stackrel{\Delta}{=} \int_0^1 P(D|p_B, G_I)P(p_B|G_I)dp_B$$
$$= B(a + c + 1, b + d + 1) = \frac{(a + c)!(b + d)!}{(N + 1)!} \tag{21}$$

$P(D|G_C)$, defined as follows, cannot be evaluated analytically, but it can be approximated by Monte Carlo method. Assuming uniform priors on p_C and p_B, it can be calculated drawing m samples of p_C and p_B from a uniform distribution on [0, 1] as follows:

$$P(D|G_C) \stackrel{\Delta}{=} \int_0^1 \int_0^1 P(D|p_C, p_B, G_C)P(p_C, p_B|G_C)dp_C dp_B$$
$$\approx \frac{1}{m}\sum_{i=1}^{m} P(D|p_{Ci}, p_{Bi}, G_C)$$
$$= \frac{1}{m}\sum_{i=1}^{m}\prod_{\mathbf{E},\mathbf{C}} P(\mathbf{E}|\mathbf{C}, B; p_{Ci}, p_{Bi})^{N(\mathbf{C},\mathbf{E})}$$
$$= \frac{1}{m}\sum_{i=1}^{m}(1 - \bar{p}_C\bar{p}_B)^a(\bar{p}_C\bar{p}_B)^b p_B^c \bar{p}_B^d, \tag{22}$$

where $\mathbf{C} = \{C, \bar{C}\}$, $\mathbf{E} = \{E, \bar{E}\}$, $N(C, E) = a$, $N(C, \bar{E}) = b$, $N(\bar{C}, E) = c$, and $N(\bar{C}, \bar{E}) = d$; and $\overline{p}_C$ and $\overline{p}_B$ indicate $1 - p_C$ and $1 - p_B$, respectively.

As this index runs from infinitely small to infinitely great, $[-\infty, \infty]$, to keep the range between –1 to 1, we define an index that is transformed by the hyperbolic tangent as an non-parameterized model based on causal support as follows:

$$\tanh\left(\frac{1}{2}SP\right) = \frac{1 - e^{-SP}}{1 + e^{-SP}}. \tag{23}$$

Note that we also tested this index as a parameterized model as Griffiths and Tenenbaum (2005) suggested in Meta-analysis 2.

A.2. Signal Detection Theory (SDT) measure

According to the parametric SDT (Green & Swets, 1966/1988; Tanner & Swets, 1954), the detectability, d', is calculated as follows:

$$d' \stackrel{\Delta}{=} \Phi^{-1}[1 - P(Y|n)] - \Phi^{-1}[1 - P(Y|s)]$$
$$= \Phi^{-1}\left(\frac{d}{c + d}\right) - \Phi^{-1}\left(\frac{b}{a + b}\right). \tag{24}$$

Here, $P(Y|n)$ is the probability of a "yes" response under noise alone, $P(Y|s)$ is the probability of a "yes" response under signal plus noise, each of which corresponds to $P(\bar{E}|\bar{C})$ and $P(\bar{E}|C)$, respectively; and Φ indicates the cumulative normal distribution function.

As this index runs from infinitely small to infinitely great, $[-\infty, \infty]$, to keep the range between –1 to 1, we adopt its transformation by the cumulative normal distribution function (i.e., so-called inverse probit transformation) as an index based on the SDT:

$$SDT \triangleq \Phi\left[\Phi^{-1}\left(\frac{d}{c+d}\right) - \Phi^{-1}\left(\frac{b}{a+b}\right)\right]. \tag{25}$$

A.3. Choice Theory measure

In his Choice Theory, Luce (1959, 1963) defined a measure of similarity, η, between signal and noise as follows:

$$\eta \triangleq \sqrt{\frac{P(N|s)}{P(Y|s)} \cdot \frac{P(Y|n)}{P(N|n)}} = \sqrt{\frac{bc}{ad}}. \tag{26}$$

According to Luce (1959, 1963), –log η is mental distance. As the mental distance between signal and noise becomes large, the accuracy of response increases, and the linkage between stimuli and responses becomes high. Thus, –log η indicates the strength of association between two events. This index, however, also runs range $[-\infty, \infty]$. Here, we define an index that is transformed by the hyperbolic tangent to keep the range between –1 to 1:

$$\tanh\left(-\frac{1}{2}\log\eta\right) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}. \tag{27}$$

This index is equivalent to Yule's (1900) coefficient of colligation, Y [25] (Equation 29).

A.4. Yule's coefficients

Yule's (1900) coefficient of association, Q, and coefficient of colligation (Yule, 1912), Y are defined as follows:

$$Q \triangleq \frac{ad - bc}{ad + bc} = \tanh\left(\frac{1}{2}\log\frac{ad}{bc}\right), \tag{28}$$

$$Y \triangleq \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} = \tanh\left(\frac{1}{4}\log\frac{ad}{bc}\right). \tag{29}$$

It is known that Y is always smaller than Q in their absolute values. Q and Y always run between –1 to 1 regardless of the marginal frequencies of the table, and this point is the greatest difference from ϕ.

A.5. Goodman–Kruskal's index of predictive association

As described in the text, Goodman and Kruskal (1954, 1963) defined an asymmetric statistic. Swapping C and E, another symmetrical index can also be defined. In terms of a 2

$\times$ 2 table, these indices are defined as follows:

$$\lambda_{E|C} \triangleq \frac{\max(a,b)+\max(c,d)-\max(a+c,b+d)}{\min(a+c,b+d)}, \tag{30}$$

$$\lambda_{C|E} \triangleq \frac{\max(a,c)+\max(b,d)-\max(a+b,c+d)}{\min(a+b,c+d)}. \tag{31}$$

As the above two indices usually do not have an equal value (i.e., $\lambda_{E|C} \neq \lambda_{C|E}$), Goodman and Kruskal (1963) defined an average index of $\lambda_{E|C}$ and $\lambda_{C|E}$, as the following:

$$\lambda \triangleq \frac{n_{E|C}+n_{C|E}}{d_{E|C}+d_{C|E}}. \tag{32}$$

Here, $n_{E|C}$ and $n_{C|E}$ indicate numerators of $\lambda_{E|C}$ and $\lambda_{C|E}$, respectively; and $d_{E|C}$ and $d_{C|E}$ indicate a denominator of $\lambda_{E|C}$ and $\lambda_{C|E}$, respectively.

A.6. Kappa statistic

In the case of a 2 $\times$ 2 table, letting p_{ij} be the probability cell ij (row i and column j), $\pi_{0,}$ the observed probability of agreement, and π_E, the probability of coincidence by chance, are expressed as follows:

$$\begin{aligned}\pi_0 &= \sum_i P_{ii} = \frac{a+d}{N},\\ \pi_E &= \sum_i p_{i+}p_{+i} = P(C)P(E)+P(\bar{C})P(\bar{E})\\ &= \frac{(a+b)(a+c)+(b+d)(c+d)}{N}.\end{aligned} \tag{33}$$

Therefore, κ it is defined as follows:

$$\kappa \triangleq \frac{\pi_0-\pi_E}{1-\pi_E} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}. \tag{34}$$

A.7. Log odds ratio

The log odds ratio of diagonal cell frequencies of a 2 $\times$ 2 contingency table is defined as follows:

$$X \triangleq \log\frac{ad}{bc}. \tag{35}$$

If you transform it by the hyperbolic tangent as it runs range [$-\infty$, ∞], it coincides with Yule's (1900) Q [24] (Equation 28) as follows:

$$\tanh\left(\frac{1}{2}X\right) = \frac{ad-bc}{ad+bc}. \tag{36}$$

A.8. Probabilistic extension of Mill's (1843/1973) method of difference

Cheng and Novick (1992) proposed a probabilistic interpretation of this idea as follows:

$$P(C|E) - P(C|\bar{E}) = \frac{a}{a+c} - \frac{b}{b+d}. \tag{37}$$

This is formally equivalent to ΔP^c [3].

A.9. Good's probabilistic causal model

Good (1961, 1962) defined a probabilistic causality as follows:

$$Q(E:C) \stackrel{\Delta}{=} \log \frac{P(\bar{E}|\bar{C})}{P(\bar{E}|C)} = \log \frac{d(a+b)}{b(c+d)}. \tag{38}$$

As this index also runs range [–∞, ∞], we define a transformed index by the hyperbolic tangent as follows:

$$G \stackrel{\Delta}{=} \tanh\left(\frac{1}{2}Q(E:C)\right) = \frac{ad-bc}{d(a+b)+b(c+d)}. \tag{39}$$

Appendix B

Model		AS95	BCC03.1	BCC03.3	LS00	W03.2	W03.6	Exp 1	Exp 2
1 H	β_0	.0289	−.140	−.873	−.809	−.0634	.0994	−.131	−.109
	β_1	.797	1.00	1.73	1.81	.879	.581	1.04	.971
35 J	x_1	.116	.411	.722	.654	.585	.169	.283	.0100
	x_2	.891	.773	.377	.579	.611	.843	.885	.443
36 ΔP_{w1}	w_1	1.23	.753	2.47	1.04	.823	.797	1.30	.000
	w_2	35.4	3.22	2.46	2.77	1.82	5.53	12.6	.203
37 ΔP_{w2}	β_0	.189	.208	−.428	−.017	.398	.156	.0566	.416
	β_1	.723	.683	1.33	1.03	.173	.220	.891	−.0251
	β_2	.274	.417	.499	.664	−.0359	.220	−.290	.0421
8 SP	β_0	47.7	43.1	13.0	31.4	47.5	1.42×10^3	45.7	20.9
	β_1	14.5	11.8	15.7	8.80	.0317	1.38×10^3	17.0	16.2
	γ	.739	.680	.600	.573	5.00	.0126	.883	.542
38 I_1	w_0	−.412	−.355	−.123	−.0144	−.0959	—	−.210	−.373
	w_1	.0408	.355	1.08	1.18	.322	—	.138	.130
	w_2	.028	.233	.307	.547	.216	—	.105	.0198
	w_3	.101	.633	.671	1.38	.238	—	.0859	.00236
39 I_2	β_0	.573	.549	.236	.499	.670	—	.508	.276
	β_1	.425	.683	1.33	1.03	.570	—	.724	2.18

(*Continued on next page*)

(*Continued*)

	Model	AS95	BCC03.1	BCC03.3	LS00	W03.2	W03.6	Exp 1	Exp 2
	β_2	.639	.683	1.33	1.03	1.10	—	.835	5.18
	β_3	.415	.833	.998	1.33	.857	—	.429	.305
40 B	p_r	.413	.360	.0584	.0523	.327	—	.351	.216
	p_c	.513	.534	.580	.0137	.491	—	.520	.772
	p_a	.416	.383	.357	.0100	.383	—	.367	.549
	p_n	.440	.440	.421	.0113	.418	—	.399	.543
41 L_*	β_0	.420	.341	−.0134	−.608	.207	—	.361	.325
	β_1	.0314	.0427	.0554	.0392	.0430	—	.0595	.0652
	β_2	−.0257	−.0427	−.0554	−.00294	−.0270	—	−.0461	−.129
	β_3	−.0141	−.0260	−.0208	−.00699	−.0164	—	−.0196	−.0104
	β_4	.00821	.0260	.0208	.0184	.0193	—	−.00982	−.000877