Causal Induction and the Revision of Belief

Daniel G. Yarlett (yarlett@psych.stanford.edu)
Michael J.A. Ramscar (michael@psych.stanford.edu)
Department of Psychology, Building 420, 450 Serra Mall,
Stanford University, Stanford, CA 94305.

Abstract

We propose that the process of causal induction can be regarded as a form of belief-revision, and formalize this idea using a discrepancy-based learning algorithm similar to that employed in the Rescorla-Wagner model of associative learning (Rescorla and Wagner, 1972) and the Belief-Adjustment Model (Hogarth and Einhorn, 1992). We then demonstrate that this model can account for conflicting patterns in human induction judgments reported by Wasserman et al. (1993) and Buehner and Cheng (1997), two data-sets which it is difficult for other models to satisfactorily explain.

Causal Induction

Most psychological work on causal induction has focused on situations in which people are presented with a series of observations concerning two factors, a candidate cause and effect, and asked to judge the degree to which they think the candidate cause is responsible for causing the candidate effect. Predictions of the degree to which a candidate cause and effect are related are usually based on the frequencies with which different configurations of a candidate cause and effect are observed to occur. With two factors, there are four possible configurations that can occur (as depicted in Table 1). These are usually taken to be the prime determinants of causal strength by such models, although other factors, such as the specific mechanism by which the causal relationship in question might be effected, have also been argued to be important (Ahn, Kalish, Medin and Gelman, 1995; Ahn and Bailenson, 1996).

Two particularly influential computational accounts of human causal induction behavior are the ΔP statistic and the Power PC model. These models make conflicting sets of predictions, and recent empirical investigations have been directed at determining which theory offers the best account of human behavior. We now describe the two theories and the characteristics of two data-sets that have proven to be pivotal in debates aimed at establishing one measure over the other.

The ΔP Statistic

ΔP is defined as the difference between the conditional probability of a candidate effect given the presence and absence of a candidate cause; it is thus sensitive to the degree to which the presence of a candidate cause *changes* the base-rate of the candidate effect occurring. The ΔP statistic is defined for a candidate cause, c, and a candidate effect, e, as below, and has been advocated by many theorists as a predictor of human judgments (Jenkins and Ward, 1965; Salmon, 1965; Rescorla, 1968; Allen, 1980; Cheng and Novick, 1990, 1991, 1992; Cheng and Holyoak, 1995):

	$+$ Effect (e)	Effect ($\bar{e}$)
$+$ Cause (c)	A	B
Cause ($\bar{c}$)	C	D

Table 1: The contingency table for a candidate cause and effect. Letters A-D represent the frequency with which each contingency is observed over a series of learning trials.

$$\Delta P = P(e|c) \quad P(e|\bar{c}) = \frac{A}{A+B} \quad \frac{C}{C+D} \qquad (1)$$

The value of ΔP ranges between 1 and -1, depending on whether the candidate cause makes the candidate effect more or less likely to be observed compared to its base-rate of occurrence. The basic predictions of the ΔP statistic have been confirmed on many occasions (Jenkins and Ward, 1965; Wasserman, Chatlosh, Neunaber, 1983; Cheng and Novick, 1990, 1991, 1992; Wasserman, Elek, Chatlosh and Baker, 1993; Lober and Shanks, 2000), and although various theories and studies have sought to undermine its validity (e.g. Cheng, 1997; Buehner and Cheng, 1997; Vallée-Tourangeau, Murphy and Drew, 1997), these studies have in turn been subjected to criticism (Lober and Shanks, 2000).

Power PC

The Power PC theory was proposed by Cheng (1997), and is based on a normative analysis of causal induction. The theory is predicated on the idea that not all candidate causes get the same opportunity to demonstrate their causal efficacy. More specifically, if we are attempting to assess the impact of a candidate generative cause on a candidate effect, then the more other generative causes interact with the candidate effect, the less chance the candidate cause has to demonstrate that it was responsible

for the occurrence of the candidate effect. An analogous argument applies to assessing preventative causes.

It can be shown that if it is assumed that multiple causes interact to produce a candidate effect according to a noisy-OR gate (Pearl, 1988), that the Power PC theory can be derived on the assumption that causal strength is assessed using maximum likelihood estimates (Glymour, 1998; Tenenbaum and Griffiths, 2000). This derivation yields the following two definitions of causal strength, one for generative causes:

$$p = \frac{\Delta P}{1 \quad P(e|\bar{c})} \tag{2}$$

And one for preventative causes:

$$q = \frac{\Delta P}{P(e|\bar{c})} \tag{3}$$

From these it can be seen that the Power PC model is essentially a 'corrected' form of ΔP and the models accordingly make distinct predictions.

An Empirical Controversy

Two data-sets have recently proven to be of pivotal significance in the debate to establish which of these two theories offers the best account of human induction judgments: the data collected by Wasserman, Elek, Chatlosh and Baker (1993), and Buehner and Cheng (1997).

Wasserman et al. (1993). The data collected by Wasserman, Elek, Chatlosh and Baker (1993) constitutes one of the most extensive investigations into human causal induction. In a series of studies they systematically varied both $P(e|c)$ and $P(e|\bar{c})$ across five discrete levels – 0.00, 0.25, 0.50, 0.75 and 1.00 – resulting in 25 experimental conditions in total. In these studies participants had to decide the degree to which pressing a telegraph key in a given time interval (1-s) affected the likelihood of a light coming on. Participants were asked to provide ratings between -100 (meaning that pressing the key always prevented the light from coming on) to +100 (meaning that pressing the key always caused the light to come on) to describe their judgments.

The results of this study highlight an important feature of human causal induction known as 'fanning' (Wasserman *et al.*, 1993; Wasserman *et al.*, 1983). Fanning can be observed when causal induction strengths with the same ΔP contingency but varying values of $P(e|\bar{c})$ are examined. In such cases, even though ΔP is constant, generative judgments get stronger as $P(e|\bar{c})$ *increases*, and preventative judgments get stronger as $P(e|\bar{c})$ *decreases*. Wasserman *et al.* (1993) describe the fanning phenomenon as follows:

> "Particularly with extreme contingency values, positive contingencies were more positively rated and negative contingencies were more negatively rated when $P(O|R) + P(O|\bar{R})$ was a small sum then when $P(O|R) + P(O|\bar{R})$ was a large sum." (p.177)

This leads to the characteristic opposed fanning patterns that can be observed in isocontingency lines. The finding is a robust one that has been found in multiple experiments (Wasserman *et al.*, 1983; Wasserman *et al.*, 1993), and which is unexplained by current models of causal induction. For example, despite the fact that the ΔP statistic correlates highly with the Wasserman *et al.* (1993) data, it systematically fails to predict any fanning, and therefore provides an incomplete description of this data.

Buehner and Cheng (1997). The studies conducted by Buehner and Cheng (1997) differ from the Wasserman *et al.* (1993) studies in two key ways. The format employed in the Wasserman *et al.* (1993) study was a response-outcome judgment task: participants were required to assess the degree to which a voluntarily initiated operation (pressing a telegraph key) had an impact on the occurrence of an observed variable (the flashing of a light). Moreover, subjects in this task were asked to assess the degree to which the occurrence of the candidate effect depends on the occurrence of the candidate cause. In the Buehner and Cheng studies, participants were not required to act so as to bring about the candidate cause, but instead merely had to observe the relationship between the two candidate variables; moreover, they were asked to assess generative and preventative strength instead of the degree to which the occurrence of the candidate effect depended on the candidate cause.

Conflicting Patterns of Judgment

Despite varying the contingencies of a candidate cause and a candidate effect across the same set of values, the results of the Wasserman *et al.* (1993) and Buehner and Cheng (1997) studies are strikingly different. In addition, neither the ΔP statistic or the Power PC model can provide adequate explanations of both data-sets. The ΔP statistic does predict the main ordinal trends in the Wasserman *et al.* (1993) data, although it fails to account for the 'fanning' effect; however, when it comes to the Buehner and Cheng (1997) data, ΔP clearly mispredicts the main trends in people's judgments. One of the most problematic aspects of the Buehner and Cheng (1997) data is that the human judgments clearly deviate from 0 when the probability of observing the candidate effect is the same in the absence as in the presence of the candidate cause (i.e. when $\Delta P = 0$). However, the Power PC model fares somewhat worse overall. It only predicts the correct trend in the Wasserman *et al.* (1993) data for negative contingencies (cases when $\Delta P < 0$), predicting the opposite trend to that actually found in the positive contingencies. Moreover, although on first inspection it appears to make correct predictions for the Buehner and Cheng (1997) data – as Buehner and Cheng originally argued – a closer inspection of the model-fit reveals that in fact it fails to predict the patterns in the data just as badly as the ΔP model (Lober and Shanks, 2000).

The direct conflict between the ratings of people in the

Wasserman *et al.* (1993) and Buehner and Cheng (1997) studies is a cause for concern if we are not to abandon the process of formally modeling these judgments – how can the same contingencies result in systematically differing ratings in the 2 studies? We believe the key to the solution lies in a closer analysis of the instructions given to participants in the two tasks. In the Wasserman *et al.* (1993) studies participants were asked to judge "whether tapping a telegraph key has any effect on the occurrence of a white light." They were asked to register their judgments on a scale ranging from +100, denoting 'causes the light to occur', to -100, denoting 'prevents the light from occurring' (p.176). In contrast, in the Buehner and Cheng (1997) studies "subjects were asked to evaluate the effectiveness of the studied vaccine at *preventing* the disease related to the virus in question" in Experiment 1A, and in Experiment 1B subjects had to give "a rating of how strongly they thought the particular rays *cause* mutation" (pp.56-57; italics added).

We wondered whether the different instructions given to participants in the two tasks could have made their evaluative responses to the same types of observation differ, hence explaining the conflicting patterns of judgment in the two data-sets. In order to test this we therefore propose a novel belief-revision framework for causal induction, and then investigate the way in which beliefs might be updated by participants as a result of the instructions they received in the different tasks.

Causal Belief-Revision Theory

Causal Belief-Revision theory defines the manner in which the degree of belief attaching to specific causal hypotheses is updated as a result of observations about a candidate cause and effect. Levels of belief are constrained to be in the range 0 to 1, where 0 indicates that the belief is not subscribed to at all, and 1 indicates that the belief is held with the highest possible conviction. The versatility of this framework is that the same mechanism can be used to explain the updating of distinct causal beliefs; we employ this idea later in explaining the conflicting patterns revealed in the Wasserman *et al.* (1993) and the Buehner and Cheng (1997) data-sets. For the minute, however, we simply describe the belief updating mechanism.

The core idea of the learning algorithm we present is that the particular observations drawn from the contingency table for a candidate cause and effect (see Table 1) either serve to provide positive or negative evidence for specific causal beliefs that are tracked because of the instructions participants are given. The idea is that each observation about a candidate cause and effect (e.g. observing the candidate cause to occur, and the candidate effect fail to occur) induces us to update our causal beliefs by a specific amount, Δb^t. This quantity is used in the following manner to yield a level of belief in a causal hypothesis at any given time step:

$$b^{t+1} = b^t + \Delta b^t \tag{4}$$

Now we need to define how Δb^t is updated. If an observation is made which counts as positive evidence for a belief, then the degree of belief is updated in the following fashion:

$$\Delta b^t = \eta W_o (1 \quad b^t)^i \tag{5}$$

And if an observation is made which counts as negative evidence for a belief, the degree of belief is updated as follows:

$$\Delta b^t = \quad \eta W_o (b^t)^i \tag{6}$$

In these equations, η is the learning rate of the model, W_o is the degree to which observation o agrees or disagrees with the causal belief, and i is a parameter which can be greater than 0, and determines the rate at which the size of discrepancies is translated into belief adjustments. At the heart of these equations is a discrepancy-based learning algorithm that bears similarities to both the Rescorla-Wagner model, and the anchoring-and-adjustment model of belief revision (especially in estimation mode; see Hogarth and Einhorn, 1992). Beliefs are updated primarily as a function of the discrepancy between the ultimate level of belief that the observed evidence should induce (this is 1 for evidence that supports the belief, and 0 for evidence that conflicts with the belief), and the current level of belief. This update scheme results in the characteristic negatively-accelerated learning curves that have been observed in a number of experimental investigations of human learning and belief revision (see Hogarth and Einhorn, 1992; Baker *et al.*, 1989; Shanks, López, Darby and Dickinson, 1996).

The associative learning framework we have specified is generally applicable to the problem of causal induction as well as the revision of beliefs in general – provided the relationship between specific observations and the beliefs in question is clear, the model can generate predictions. For example, it can readily be used to compute the ΔP statistic – just as the Rescorla-Wagner model can (see Chapman and Robbins, 1990; also Cheng, 1997) – under minimal assumptions. However, before the model can be applied to the current data-sets, it is necessary to determine how the causal beliefs that are assessed in the two studies are updated in the light of the four possible observations possible with two binary factors.

Study 1

The aim of this study was to find out how the causal beliefs that participants were assessing in the Wasserman *et al.* (1993) and Buehner and Cheng (1997) studies would be updated as a result of observations in the experiments, and consequently whether instructional differences could be responsible for the conflicting patterns of judgments. Participants were therefore provided with the salient aspects of the instructions to one of the 2 original studies, and then asked to rate to what degree they thought each possible observation (as set out in Table 1) agreed or disagreed with the causal belief the instructions asked them to assess.

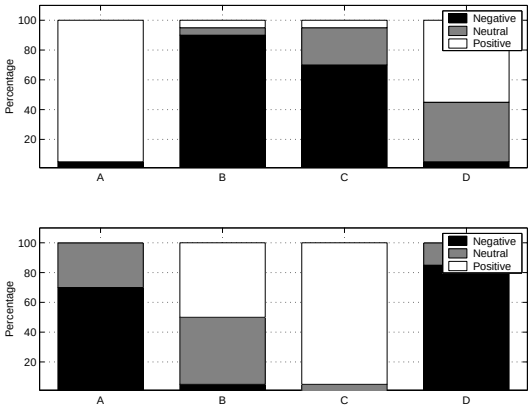

Figure 1: Results from the Wasserman *et al.* (1993) condition.

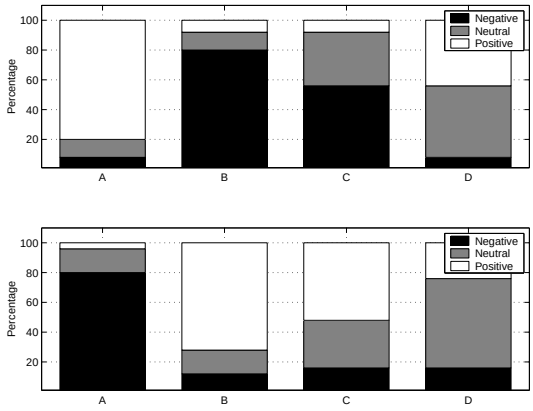

Figure 2: Results from the Buehner and Cheng (1997) condition.

Methods and Design. Participants were randomly assigned to one of 2 conditions in which they were either asked about the Wasserman *et al.* (1993) or the Buehner and Cheng (1997) study. Depending on what condition they were assigned to, participants were either asked to imagine that they were assessing (i) the degree to which a factor B occurs in the presence and absence of another factor A, or (ii) the degree to which a factor A generates or prevents another factor B. Once participants had been provided with this information, for each of the two beliefs they had been asked to track, they were presented with the 4 possible observations that could occur (see Table 1) in a random order, and asked to rate how much they thought the observation 'agreed' or 'disagreed' with the belief that they were tracking. Ratings were collected on a 9-point Likert scale, in which 1 meant that the observation strongly disagreed with the causal belief; 5 meant that the observation was irrelevant to assessing the belief; and 9 meant that the observation strongly agreed with the belief.

Participants. 45 students at the University of Edinburgh, Scotland, participated in the study. All participants were volunteers, and no reward was offered for taking part in the study.

Results. Because we were chiefly interested not in the *strength* which certain observations were judged to agree or disagree with the beliefs about the candidate factors, but instead the *valency* of the evidence they provide, the ratings were first grouped into three categories: Positive, Negative and Neutral. Ratings were recorded as Positive if greater than or equal to 6; Neutral if equal to 5; and Negative if less than or equal to 4. The percentage breakdown of each of the three categories for each observation is shown in Figures 1 and 2. In order to set the parameters of our model we coded a particular observation as being Neutral for the causal hypothesis in question if 40% or more of participants had judged it as such. Otherwise each observation was coded as being Positive or Negative depending on which category received the greatest number of votes.

Discussion

The results from this study allowed us to set the W_o parameters in the Causal Belief-Revision model for the Wasserman *et al.* (1993) and Buehner and Cheng (1997) studies respectively. Interestingly, the valency judgments were very similar when participants were asked to assess the degree to which a candidate effect occurs in the presence of a candidate cause and the degree to which a candidate cause generates a candidate effect (top panels, respectively, of Figures 1 and 2). However, the judgments were divergent when participants were asked to assess the degree to which a candidate effect occurred in the absence of a candidate cause, and the degree to which a candidate cause prevents a candidate effect(bottom panels, respectively, of Figures 1 and 2), indicating the evaluative differences induced by instructional characteristics could be a potential explanation of the divergent patterns of induction judgments in the two original studies.

	A	B	C	D
Wasserman *et al.* (1993)				
Cause present	1	-1	-1	0
Cause absent	-1	0	1	-1
Buehner and Cheng (1997)				
Generative causes	1	-1	-1	0
Preventative causes	-1	1	1	0

Table 2: The values of the W_o parameters as determined in Study 1.

Simulation 1: Wasserman *et al.* Data

We used the parameters acquired in Study 1 to fit the Causal Belief-Revision model to the Wasserman *et al.* (1993) data-set. We also measured the performance of various versions of the Rescorla-Wagner model on this data-set because it has successfully been used to explain a diverse range of empirical data on induction, and therefore provides a reasonable point of comparison.

Because people were asked to assess the degree to which the candidate effect depended on the candidate cause in the Wasserman *et al.* (1993) study, this focused people on the frequency with which the candidate cause occurred in the presence and in the absence of the candidate cause. Accordingly, the degree of predicted depen-

	R^2	r_s
Causal Belief-Revision Model	0.9832	0.9910
Unrestricted Rescorla-Wagner	0.9808	0.9905
Restricted Rescorla-Wagner	0.9639	0.9831

Table 3: Empirical performance of the models on the Wasserman *et al.* (1993) data-set.

dence was defined to be:

$$b_d = b_p \quad b_a \tag{7}$$

where b_p is the belief that the candidate effect occurs when the candidate effect is present, b_a is the belief that the candidate effect occurs when the candidate effect is absent, and the levels of both beliefs are adjusted according to the Causal Belief-Revision model.

Method. Because the Causal Belief-Revision model relied on randomly generated sequences of observations, its predictions varied slightly from run to run. In order to minimize the possibility of this biasing the results, the mean performance of the model was taken over 250 trials. In addition, the mean of the 10 best parameterizations was used to define the overall performance of each model, to avoid the possibility that noise in the simulation data resulted in spurious model performance. The i parameter was fit to the data by varying it in the range 1-3 in steps of 0.1. The W_o parameters were set as shown in Table 2, according to the results of Study 1.

The predictions of each model were linearly scaled so that they were expressed across the same range as the empirical data. In order to assess the overall quantitative fit of each model we used the coefficient of determination (R^2) to measure the amount of variance in the empirical data explained by each model. However, to ensure that the overall qualitative match was not sacrificed in order to produce a better quantitative match, the Spearman's Rank correlation was also calculated for the each model with the same parameter values that produced the greatest value of R^2.

Results. The results of this Simulation are reported in Table 3. Notably the Causal Belief-Revision model outperformed the Rescorla-Wagner family of models on both qualitative and quantitative measures of fit, and overall explained an extremely high amount of the variance in the judgments (98.3%). It did this with parameters of $\eta = 0.375$ and $i = 1.54$. Moreover, this success was not accomplished simply through the addition of free parameters: the Unrestricted Rescora-Wagner model and the Causal Belief-Revision model both had 2 free parameters. These results demonstrate the success of the Causal Belief-Revision model in predicting patterns in human causal induction, in particular its ability to reproduce the 'fanning' phenomenon. The predictions of the model can be seen plotted against the Wasserman *et al.* (1993) data in Figure 3.

Discussion

The Causal Belief-Revision model successfully explains an exhaustive and representative data-set concerning human induction, and in doing so it outperforms another competitive pair of models based on the Rescorla-Wagner learning rule. Moreover, although both sets of models managed to capture large proportions of variance in the empirical data, only the Causal Belief-Revision model was able to predict the qualitative fanning trends

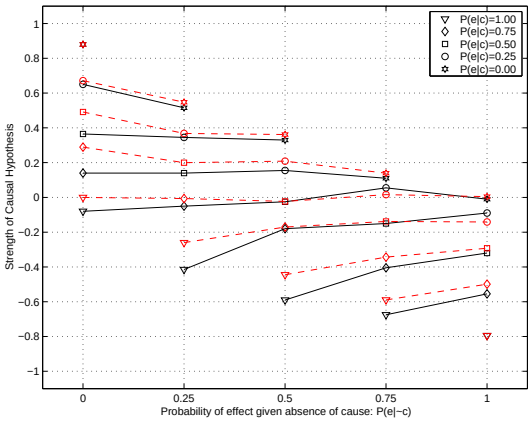

Figure 3: Predictions of Causal Belief-Revision model predicted against human performance in the Wasserman *et al.* (1993) studies (human judgments plotted in continuous line; model predictions plotted in dotted line).

actually observed. In order to further test the explanatory power of the Causal Belief-Revision framework we now examine whether it can *also* successfully explain the conflicting patterns identified in the Buehner and Cheng (1997) data-set.

Simulation 2: Buehner & Cheng Data

The Causal Belief-Revision model was fit to the Buehner and Cheng (1997) data, again along with versions of the Rescorla-Wagner model. This time the Causal Belief-Revision model was used to assess generative and preventative hypotheses across the contingencies explored by Buehner and Cheng (1997), using the valency judgments acquired in Study 1. This application of the model differed from its application to the Wasserman *et al.* (1993) in that it was only used to assess the degree of belief in one causal hypothesis at a time, instead of two.

Method. The Causal Belief-Revision model was used to make predictions across runs consisting of 16 observations, just as in the original Buehner and Cheng (1997) experiments. In order to correct for variance in the model predictions resulting from the randomly generated sequences of observations, the predictions of the model over these runs were averaged over 250 trials. Other aspects of the method were as reported for Simulation 1. The W_o parameters were set as shown in Table 2, according to the results of Study 1.

Results. The results of fitting the various models to the data are reported in Table 4. The predictions of the best-performing parameterization of the Causal Belief-Revision Model are shown plotted against the human data in Figure 4.

Discussion

The Causal Belief-Revision model produced by far the best quantitative fit to the Buehner and Cheng (1997) data. That it did so without sacrificing its representation of the qualitative trends in the data is apparent from both the high Spearman's Rank correlations ($r_s = 0.97$ for

	R^2	r_s
Causal Belief-Revision Model	0.9689	0.9679
Causal Belief-Revision Model	0.9783	0.9750
Restricted Rescorla-Wagner	0.6397	0.7691
Restricted Rescorla-Wagner	0.7904	0.8834
Unrestricted Rescorla-Wagner	0.8091	0.5746
Unrestricted Rescorla-Wagner	0.8500	0.7601

Table 4: Summary empirical performance of the models on the Buehner and Cheng (1997) data-set. Performance for preventative (Experiment 1A), then generative (Experiment 1B), judgments are given for each model.

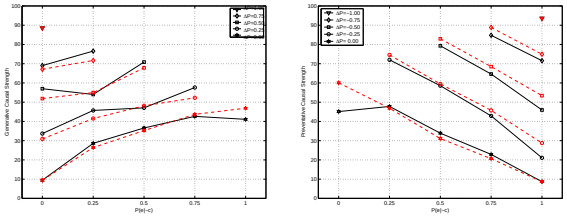

Figure 4: Performance of the Causal Belief-Revision model plotted against human performance in the Buehner and Cheng (1997) studies (left panel for generative judgments; right panel for preventative judgments). Human data plotted in continuous lines, model predictions in dotted lines.

the preventative data, and $r_s = 0.98$ for the generative data) and the plot of the empirical against the predicted results (Figure 4). Crucially, we have demonstrated that the Causal Belief-Revision model can provide explanations of the conflicting patterns in *both* the Wasserman *et al.* (1993) and the Buehner and Cheng (1997) data.

General Discussion

We have presented a novel model of human causal induction judgments, the Causal Belief-Revision theory, and shown that it is capable of reconciling the conflicting patterns in two key empirical data-sets. To our knowledge no other current model of causal induction is able to provide an explanation of both these data-sets in a satisfactory way, and indeed it is difficult to see how a model could do so without having recourse to the idea of instructional differences affecting evaluative responses that we have introduced. We believe that the Causal Belief-Revision model provides a new way of conceptualizing causal induction behavior, and intend to pursue this approach further in the future.

Acknowledgments

The authors thank Gordon Bower, Nicolas Davidenko, Danny Oppenheimer and Ewart Thomas for their insightful comments on this work, and also Marc Buehner for providing the Buehner and Cheng (1997) data.

References

Ahn W. and Bailenson J. (1996). Causal Attribution as a Search for Underlying Mechanisms: An Explanation of the Conjunction Fallacy and the Discounting Principle, *Cognitive Psychology*, **31**, 82-123.

Ahn W., Kalish C.W., Medin, D.L. and Gelman S.A. (1995). The Role of Covariation Vs. Mechanism Information in Causal Attribution, *Cognition*, **54**, 299-352.

Baker A.G., Berbier M.W. and Vallee-Tourangeau F. (1989). Judgements of a 2×2 Contingency Table: Sequential Processing and the Learning Curve, *The Quarterly Journal of Experimental Psychology*, **41B**(1), 65-97.

Buehner M. J. and Cheng P.W. (1997). Causal induction: The Power PC Theory Versus the Rescorla-Wagner Model. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 55-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chapman G.B. and Robbins S.J. (1990). Cue Interaction in Human Contingency Judgment, Memory & Cognition, 18, 537-545.

Cheng P. (1997). From Covariation to Causation: A Causal Power Theory, *Psychological Review*, **104**, 367-405.

Cheng P.W. and Novick L.R. (1990). A Probabilistic Contrast Model of Causal Induction, *Journal of Personality and Social Psychology*, **58**, 545-567.

Cheng P.W. and Novick L.R. (1991). Causes Versus Enabling Conditions, *Cognition*, **40**, 83-120.

Cheng P.W. and Novick L.R. (1992). Covariation in Natural Causal Induction, *Psychological Review*, **99**, 365-382.

Glymour C. (1998). Learning Causes: Psychological Explanations of Causal Explanations, *Minds & Machines*, **8**(1), 39-60.

Hogarth R.M. and Einhorn H.J. (1992). Order Effects in Belief Updating: The Belief-Adjustment Model, *Cognitive Psychology*, **24**, 1-55.

Lober K. and Shanks D.R. (2000). Is Causal Induction Based on Causal Power? Critique of Cheng (1997), *Psychological Review*, **104**(1), 195-212.

Pearl J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers, Palo Alto.

Rescorla R.A. and Wagner A.R. (1972). A Theory of Pavlovian Conditioning: Variations on the Effectiveness of Reinforcement and Non-reinforcement. In A.H. Black & W.F. Prokasy (Eds.), Classical Conditioning II: Current Research and Theory (pp.64-99). New York: Appleton-Century-Crofts.

Tenenbaum J.B. and Griffiths T.L. (2000). Structure Learning in Human Causal Induction, *Advances in Neural Information Processing Systems*, **13**, MIT Press, Cambridge, MA.

Ward W.C. and Jenkins H.M. (1965). The Display of Information and the Judgment of Contingency, *Canadian Journal of Psychology*, **19**, 231-241.

Wasserman E.A., Chatlosh D.L. and Neunaber D.J. (1983). Perception of Causal Relations in Humans: Factors Affecting Judgments of Response-Outcome Contingencies Under Free-Operant Procedures, *Learning and Motivation*, **14**, 406-432.

Wasserman E.A., Elek S.M., Chatlosh D.L. and Baker A.G. (1993). Rating Causal Relations: Role of Probability in Judgments of Response-Outcome Contingency, *Journal of Experimental Psychology: Learning, Memory and Cognition*, **19**(1), 174-188.