The Role of Reliability in Human Contingency Judgment

Matthew A. Jacques (mjacq012@uottawa.ca)
School of Psychology, University of Ottawa
145 Jean-Jacques-Lussier Street, Ottawa, Ontario K1N 6N5

Pierre Mercier (pierre.mercier@uottawa.ca)
School of Psychology, University of Ottawa
145 Jean-Jacques-Lussier Street, Ottawa, Ontario K1N 6N5

Abstract

This report describes an experimental investigation into the role of reliability in contingency judgment. Cheng's (1997) PowerPC theory of human contingency judgment was designed to address the inadequacies of both earlier associative (Rescorla & Wagner, 1972) and computational (Jenkins & Ward, 1965) models. These earlier models were argued to be restrictive in that they deal solely with co-variation, while the PowerPC theory incorporates the aspect of causal power (Cheng, 1997). Prior tests of the PowerPC theory have returned inconsistent results (Lober & Shanks, 2000; Vallée-Tourangeau, Murphy & Drew, 1997). Proponents of PowerPC maintain that such results can be accounted for by their theory if reliability is taken into consideration (Beuhner, Cheng, & Clifford, 2003; Buehner & Cheng, 1997). This concept of reliability as an adjunct to the original Power PC theory is discussed, and it is demonstrated both logically and computationally that such a revision of the theory would be tantamount to a return to the earliest co-variation model which the PowerPC was intended to improve upon.

Keywords: contingency, judgment, causal, power, reliability.

Background

It is possible to summarize contingencies as 2x2 tables, which represent a cross-tabulation of the presence or absence of two events. When a contingency is hypothesized to have some causal mechanism, one of these events is said to be the cause while the other is the effect. As illustrated below, each of the four cells in the 2x2 contingency table accounts for one of the possible combinations - Cause Present and Effect Present (*a*), Cause Present and Effect Absent (*b*), Cause Absent and Effect Present (*c*), and Cause Absent and Effect Absent (*d*).

Table 1: A simple 2 x 2 contingency table.

	Effect present	Effect absent
Cause present	*a*	*b*
Cause absent	*c*	*d*

Jenkins and Ward (1965) first proposed that this relationship or contingency between the cause and the effect in this table can be quantitatively expressed as the Δ*P* coefficient. In this model, contingency is calculated as the difference between the probability of the effect in the presence of the cause minus the probability of the effect in the absence of the cause. The ΔP calculation applies to contingencies between two binary events. This generates the metric in Equation 1, expressed as cell frequencies, or as in Equation 2 expressed in terms of probability where *e* represents the effect, and *i* represents the candidate cause being evaluated. And *P*(*e*|*i*) then indicates the probability of the effect given the presence of the cause, while *P*(*e*|~*i*) is the probability of the effect in the absence of the cause. At times this report will make reference to the *base rate* of the effect, an expression used interchangeably with *P*(*e*|~*i*) for the sake of simplicity as they both refer to the likelihood of the effect in the absence of any influence of the hypothesized cause being evaluated (*i*).

$$\Delta P = \frac{a}{a+b} - \frac{c}{c+d} \qquad (1)$$

$$\Delta P = P(e \mid i) - P(e \mid \sim i) \qquad (2)$$

Following these formulations, it is possible to compute both positive and negative values of Δ*P*. A positive contingency would be calculated when the effect is more likely to occur in the presence of the cause. When the hypothesized causal mechanism is intended to increase the likelihood of the effect, this is also called a *generative* contingency. A negative contingency will occur when the effect occurs less often when the cause is present. This can also be considered a *preventive* contingency, if the hypothesized mechanism behind the contingency is said to reduce the occurrence of the effect.

Competing Models of Contingency and Causality

Most models of human contingency judgment fall into two major categories – computational or associative. Purely associative models find their roots in Pavlovian classical conditioning. They describe a learning curve of increasing associative strength between a cause and effect, on a trial-by-trial basis. The Rescorla-Wagner model (1972) is a primary exemplar of this approach, and describes the change in associative strength between the cause and effect over repeated trials. According to the theory, participants do not compute the value of Δ*P* as described in equations 1 and 2, but rather are conditioned by the increasing association between cause and effect. This does well to explain the associative learning curve, but has been criticized for relying on seemingly nebulous values for cause and effect salience (Cheng, 1997). The Rescorla-Wagner model on its own, also fails to provide asymptotic, final predictions for associative strength which would allow it to be readily compared to other models' predictions.

Computational models follow the basic assumption that people will base their judgments of contingency upon mental comparisons of two likelihoods: that of a given effect in the presence or absence of a hypothesized 'candidate' cause. The value calculated by making such a comparison is expected to serve as a normative guidepost which participants will estimate by observation in the experimental setting. The Probabilistic Contrast Model (PCM) (Cheng & Novick, 1990) drew upon the ΔP of Jenkins and Ward (1965) and provided an early computational formulation which, in its simplest form, is identical to Equations 1 and 2 above. In practice, according to this model, individuals' judgments of contingencies are expected to follow *ordinally* from the normative values calculated for ΔP. This is to say that while internally generated subjective judgments are not hypothesized to match the *exact* calculated values of ΔP, it is expected that the pattern of judgments would follow the same *order* as the normative ΔP values.

While there is a history of empirical support for most of the predictions of the PCM (Shanks, 1985), the model fails to accommodate certain conditions where either the effect occurs all the time or the effect never occurs (regardless of the presence or absence of the candidate cause). In both of these situations, ΔP is calculated to be zero due to there being *no difference* between the likelihood of the effect in the presence or absence of the cause. However, a prediction of no effect of the candidate cause in both of these situations could be erroneous, and research participants do not normally have difficulty responding to questions about such conditions – indicating that we are still able to make decisions even in such relatively uncertain or ambiguous circumstances (Cheng, 1997).

Cheng's (1997) Power PC Theory

To address some of the shortcomings of the PCM, Cheng's (1997, 2000) Power PC theory strives to account for both of these problematic conditions (effect always occurring and effect never occurring) with one causal power theory of contingency. The concept of causal power was introduced as an improvement over purely co-variational models. In effect, it states that causal reasoners "…induce the unobservable *causal power* of a candidate cause in the distal world from observable events represented in the proximal stimulus." (Buehner, Cheng & Clifford, 2003, p. 1120). The computational equation of the Power PC theory for generative causes is:

$$p_i = \frac{\Delta P}{1 - p(e \mid \sim i)} \tag{3}$$

In Equation 3, ΔP represents the level of contingency between the cause and the effect, *P(e|~i)* is the base rate of the effect, and p_i is the predicted level of causal power of the candidate cause *(i)*. Several of the assumptions of the Power PC theory relate to how the candidate cause *(i)* must act independently of all alternate causes of the effect *(a)* :

> "(1) When *e* occurs, it produces *e* with probability p_i; when *a* occurs, it produces *e* with probability p_a; and nothing else influences the occurrence of *e*;
> (2) *i* and *a* influence the occurrence of *e* independently; and
> (3) *i* and *a* influence the occurrence of *e* with causal powers that are independent of how often *i* and *a* occur"
> (Cheng1997, p. 373)

Further, Cheng (1997) provides a separate formulation of the Power PC theory for preventive contingencies. This occurs when the calculated value of ΔP is negative, due to the likelihood of $p(e|{\sim}i)$ exceeding that of $p(e|i)$. In such cases, causal power is instead calculated using the following formula:

$$p_i = \frac{-\Delta P}{p(e \mid \sim i)} \tag{4}$$

The interplay of the first two components that make up the Power PC formula is illustrated in Table 2 for both positive (generative, using Equation 3) and negative (preventive, using Equation 4) contingencies. For any given value of ΔP (left side of table), with a specified base rate of *P(e|~i)* (along the top of table), the predicted causal power levels are provided.

As shown in the Table 2, at somewhat extreme values of low contingency (i.e., $\Delta P = .20$) and high base rate (i.e., P(e|~i) = .8), the prediction of the Power PC theory (p = 1*)* appears implausible. Intuitively, it seems unlikely that participants would judge conditions of low contingency as having maximal causal power. Why would an event that is not systematically followed by a certain effect be judged to strongly cause that effect, even when the effect occurs frequently on its own (high base rate)? There is some previous research which has examined p_i as a function of base rate. For instance, Buehner and Cheng (1997), Vallée-Tourangeau, Murpy and Drew (1997), and Lober and Shanks (2000) have indeed shown a relation between *P(e|~i)* and causal power.

However, they have not comprehensively explored the many predictions contained in Table 2. With the more counter-intuitive predictions untested, relatively little is known about how the predictions of the theory hold up as ΔP approaches 0 and as the base rate remains high, as well as when ΔP approaches 1 and as the base rate remains low. In both of these type of situations, it seems puzzling that Power PC predicts maximal contingency judgments, yet if empirical support can be generated for those predictions, the theory would be confirmed as a significant improvement upon earlier associative and computational approaches.

Table 2: Matrix of Causal Power, Covariation and Base-Rate

	P(e|~i)										
Δ*P*	**0.00**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**	**1.00**
1.00	1.00										
0.90	0.90	1.00									
0.80	0.80	0.89	1.00								
0.70	0.70	0.78	0.88	1.00							
0.60	0.60	0.67	0.75	0.86	1.00						
0.50	0.50	0.56	0.63	0.71	0.83	1.00					
0.40	0.40	0.44	0.50	0.57	0.67	0.80	1.00				
0.30	0.30	0.33	0.38	0.43	0.50	0.60	0.75	1.00			
0.20	0.20	0.22	0.25	0.29	0.33	0.40	0.50	0.67	1.00		
0.10	0.10	0.11	0.13	0.14	0.17	0.20	0.25	0.33	0.50	1.00	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	N/A
-0.10		1.00	0.50	0.33	0.25	0.20	0.17	0.14	0.13	0.11	0.10
-0.20			1.00	0.67	0.50	0.40	0.33	0.29	0.25	0.22	0.20
-0.30				1.00	0.75	0.60	0.50	0.43	0.38	0.33	0.30
-0.40					1.00	0.80	0.67	0.57	0.50	0.44	0.40
-0.50						1.00	0.83	0.71	0.63	0.56	0.50
-0.60							1.00	0.86	0.75	0.67	0.60
-0.70								1.00	0.88	0.78	0.70
-0.80									1.00	0.89	0.80
-0.90										1.00	0.90
-1.00											1.00

Confidence and Reliability

Confidence is an important secondary issue in theorizing about causality judgements for two reasons. First, confidence plays an important role in any decision we make. We are intuitively less likely to make an extreme decision if we are uncertain of the evidence supporting the decision. The extent of the role that confidence plays in causal attribution remains an unknown. Second, it has been firmly established empirically that causality judgements progress gradually toward their asymptotic value (Vallée-Tourangeau, Murphy & Drew, 1997; Lober & Shanks, 2000); yet Power PC theory cannot intrinsically account for this effect without recourse to an external process such as confidence. Recall that computational models provide an asymptotic end-point prediction, compared to the learning-curve described by associative accounts. Collecting confidence data will be helpful in determining its role in the contingency judgment task, as well as providing secondary data for differentiating the predictions of associative and computational models.

It is also in relation to confidence that some researchers (Buehner & Cheng, 1997; Buehner, Cheng & Clifford, 2003) have raised the issue of how results otherwise seen as inconsistent with the Power PC theory could be due to participants confounding causal power with the concept of reliability. They define reliability as the number of opportunities the candidate cause (*i*) has to demonstrate its' effect (*e*), minus the number of these trials which the participant might expect to happen simply by chance considering the base rate of the effect (*P(e|~i)*). While the exact definition of reliability is not provided in these sources (1997, 2003), M. Buehner (personal communication, October 5, 2005) has confirmed that the formulation is as outlined in Equation 5 below.

$$R = (a+b) - (a+b)\left(\frac{c}{c+d}\right) \qquad (5)$$

Equation 5 depicts reliability as the total number of trials where the presumed cause is present (*a* + *b*), minus the proportion of those trials which could be accounted for by the base rate occurrence of the effect in the absence of the presumed cause. It is intended to be act as an adjustment or moderating factor within the scope of the Power PC theory. We have derived Equation 6 as a means of expressing reliability as a proportion rather than a raw number, by dividing the number of trials from Equation 5 by the total number of trials in which the candidate cause is present (*a* + *b*). This converts an absolute value (number of trials) into a relative one (proportion of trials) which can be controlled for across different experimental conditions.

$$R = \frac{(a+b) - (a+b)\left(\frac{c}{c+d}\right)}{(a+b)} \qquad (6)$$

It is hoped that this initial investigation will increase understanding about the role of reliability with respect to both judgment and confidence. If participants are confounding reliability with causal power, one would expect their confidence ratings to drop when reliability is low.

Experiment

Participants

Twenty-four volunteer undergraduate students (12 Male, 12 Female) from the University of Ottawa took part in the experiment. All participants were University of Ottawa undergraduate students recruited on a volunteer basis. There was no financial or academic reward for their participation.

Apparatus

The experimental task was carried out on an IBM-PC compatible desktop computer, equipped with a 17" SVGA colour monitor in an individual testing room. Data were recorded anonymously on the computer's hard drive and backed up on a 3.5" diskette after each testing session. The experimental task was programmed with Microsoft Visual Basic Professional 6.0.

Stimuli

The stimuli consisted of a graphical representation of fictional patients' medical records. These records contained either the presence or absence of a pill, and the presence or absence of a skin discoloration. The pill's presence was indicated by the presentation of either a red or blue pill, while the pill's absence was indicated with a black 'X' overlaying the icon. The skin discoloration was presented as a face the same colour as the pill, with the tongue sticking out. The absence of the skin discoloration was presented as a white, happy face. The treatment (pill) and outcome (skin condition) were presented simultaneously, for the duration of 1 second, with a delay of 1 second before the presentation of the next medical record.

Procedure

Participants were seated in front of a computer and oriented to the experiment with written instructions which appeared on the computer's screen. The initial instructions were:

"Welcome to the Cognitive Psychology Laboratory. We are studying decision-making processes such as those used by health professionals. Imagine that you have access to the medical files of many people sick with different diseases. These diseases have many symptoms, including one in common: a skin discoloration. Different medications treating these diseases have been used in clinical trials. The problem is that the medications themselves can also increase or decrease the risk of skin discoloration. We will ask you to evaluate different medications. For each medication, the clinical trial contains the results of many individual files. Each file may report the absence of medication, or the presence of medication, the presence of skin discoloration, or the absence of skin discoloration, in different possible combinations. You will have an opportunity to warm up your diagnostic skills with three practice trials."

Three practice conditions then followed, consisting of 40 medical files each. The first practice had a *ΔP* of .67, the second -.67 and the third was 0. Each experimental condition, as outlined in Table 3, also consisted of a set of 40 fictitious medical files, each of which constituted a trial. The order of presentation of the trials, and of the experimental conditions, was randomized. At the end of each condition, participants were asked to judge the relative efficacy of the treatment on a scale from -100 ("extreme reduction of skin discoloration") to 100 ("extreme increase in the risk of skin discoloration") with 0 labelled as "unchanged risk of discoloration". After making this judgment, the participants were asked to rate their confidence in the judgment they just made, on a scale from 0 ("No confidence") to 100 ("High confidence") with 50 labelled as "Moderate confidence". It was emphasized that there were no real 'right' or 'wrong' judgments, and that participants should feel free to record their judgment whether they felt they were 'just guessing' or not.

In order to test the nature of the relationship between reliability, power and causal judgments, conditions were created in which predictions of the Power PC theory, the Power PC+Reliabilty theory, and the basic PCM would be tested against each other. The same basic design was repeated with three different levels of the probability of the cause, in order to observe its effect, and rule it out as a potential confound. A 3x2x2 (Low-Med-High probability of the cause, by Low-High power, and Low-High reliability within each of those conditions) factorial design was used.

Table 3: Experimental Design

a	*b*	*c*	*d*	*P(e|i)*	*P(e|~i)*	*ΔP*	*pi*	*R*	*ProbC*
9	1	27	3	0.90	0.90	0.00	0.00	0.10	0.25
0	10	0	30	0.00	0.00	0.00	0.00	1.00	0.25
10	0	27	3	1.00	0.90	0.10	1.00	0.10	0.25
10	0	0	30	1.00	0.00	1.00	1.00	1.00	0.25
18	2	18	2	0.90	0.90	0.00	0.00	0.10	0.50
0	20	0	20	0.00	0.00	0.00	0.00	1.00	0.50
20	0	18	2	1.00	0.90	0.10	1.00	0.10	0.50
20	0	0	20	1.00	0.00	1.00	1.00	1.00	0.50
27	3	9	1	0.90	0.90	0.00	0.00	0.10	0.75
0	30	0	10	0.00	0.00	0.00	0.00	1.00	0.75
30	0	9	1	1.00	0.90	0.10	1.00	0.10	0.75
30	0	0	10	1.00	0.00	1.00	1.00	1.00	0.75

Results

The contingency judgment results presented in Figure 1 show that judgments were simultaneously influenced by power, reliability and probability of cause. When reliability was low (.1), powers of 0 vs. 1.0 could not be discriminated, except perhaps as the probability of cause increased towards .75. When reliability was high (1.0), powers of 0 vs. 1.0 could easily be discriminated and perhaps even more so as the probability of cause progressed towards .75. The influence of probability of cause is such that a power/contingency of 0 is perceived as more negative as the probability of cause increases. Causal power was judged maximal when the Power PC predicted value was 1.0, and ΔP was equal to 1.0 as well. For conditions where p = 1.0 and ΔP = .10, the observed reduction in participant judgments could arguably be explained by the reliability-adjusted Power PC theory.

These initial observations are corroborated by the statistical analyses that found a significant main effect of power ($F_{1,23}$ = 111.430) and a significant interaction of power and probability of cause ($F_{2,22}$ = 5.446). Power also interacts significantly with reliability ($F_{1,23}$ = 105.932). Finally, there was also a significant interaction between probability of cause and reliability ($F_{2,22}$ = 4.423).

Confidence results, as depicted in Figure 2, reveal that confidence is influenced by power (F1,23 = 15.256), reliability (F1,23 = 43.699) and the probability of cause (F3,46 = 3.692), although the latter two interact (F2,46 = 7.204) such that confidence is less reduced by low reliability at higher levels of the probability of cause. This lends credence to the argument that participants' confidence is indeed related to reliability. All of these effects or interactions were significant at or beyond the p = .05 level of statistical significance.

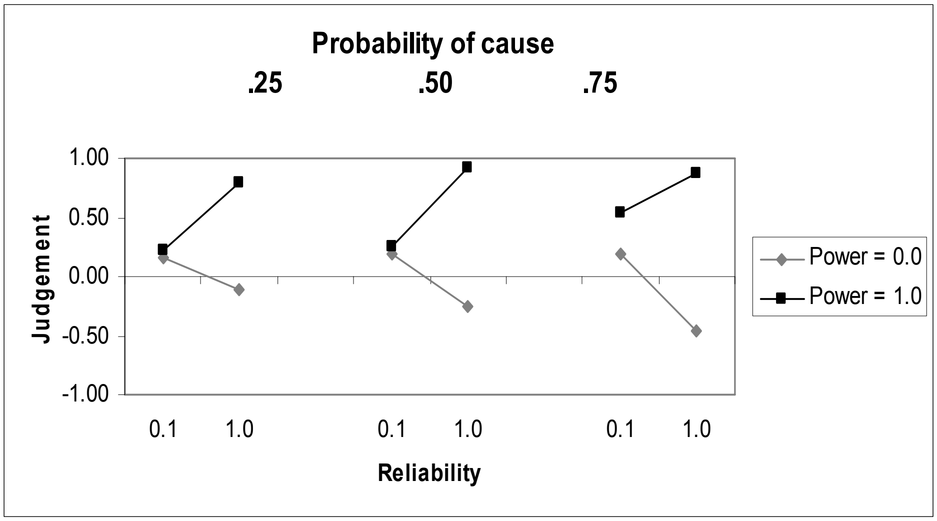

Figure 1: Judgment Results

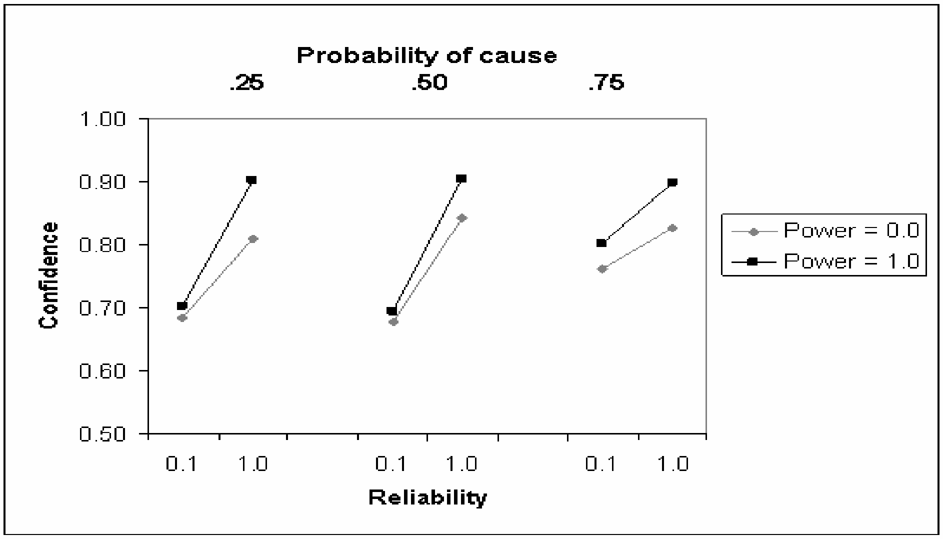

Figure 2: Confidence Ratings

Discussion

This experiment set out to explore and introduce experimental control over the concept of reliability, in order to answer the question of whether participants' judgments that deviate from p_i could be explained by reliability. In designing this experiment, no fully factorial design was found in which all three of reliability, power, and ΔP could be controlled. In the end, a main effect of reliability as well as an interaction of reliability and p_i were found. This at first appears to support the argument that an adjusted Power PC theory, factoring in reliability, could account for participants' judgments which otherwise fall short of the Power PC-predicted maximal value of 1.0. Indeed, the discovery of a statistical interaction between reliability and p_i seems to underline Beuhner and Cheng's (1997) and Buehner's (2003) conception of how reliability and p_i work in conjunction.

Returning to the difficulty producing a fully factorial design, the apparent problematic relationship between reliability, power, and ΔP was more fully explored following the experimental results. If, as Buehner and Cheng (1997) and Buehner, Cheng and Clifford (2003) hypothesize, research participants are indeed susceptible to conflating reliability with causal power, it seems rational that one would expect their judgment ratings to be predicted by the formulation outlined from Equations 7 and 8, where the new term Power PC' represents the reliability-adjusted Power PC model.

$$\textit{Power PC'} = p_i{*}R \tag{7}$$

$$\textit{Power PC'} = \left(\frac{\frac{a}{a+b}-\frac{c}{c+d}}{1-\frac{c}{c+d}}\right)\left(\frac{(a+b)-(a+b)\left(\frac{c}{c+d}\right)}{(a+b)}\right) \tag{8}$$

Exploring this interaction further, the reliability term (when expressed as a proportion) can be reduced to 1-*P(e|~i)*, resulting in Equation 9, which in turn is demonstrated to ultimately reduce to ΔP in Equations 10 through 12. Equation 11 is equivalent to the formulation of ΔP presented in Equation 1.

$$\textit{Power PC'} = \left(\frac{\frac{a}{a+b}-\frac{c}{c+d}}{1-\frac{c}{c+d}}\right)\left(1-\frac{c}{c+d}\right) \tag{9}$$

$$\textit{Power PC'} = \left(\frac{\frac{a}{a+b}-\frac{c}{c+d}}{1-\frac{c}{c+d}}\right)\left(\frac{1-\frac{c}{c+d}}{1}\right) \tag{10}$$

$$\textit{Power PC'} = \left(\frac{a}{a+b}-\frac{c}{c+d}\right) \tag{11}$$

$$\textit{Power PC'} = \Delta P \tag{12}$$

Given this finding, it is not surprising that a fully factorial design was elusive. For any given combination of causal power (p) and reliability (R), ΔP will be fixed at a determined value. Given a simple 2x2 contingency, there are only so many 'degrees of freedom', and combining causal power and reliability results in a return to the original computational formulation of Jenkins and Ward (1965).

This observation should not be taken to indicate that reliability itself is a fallible or redundant concept. On the contrary, the experimental results presented here suggest a significant role for reliability with regards to both judgments and confidence. Further research into the concept of reliability may provide a fruitful addition to our understanding of contingency judgment. The caveat presented here is that within the scope of the Power PC theory, reliability does not provide a sufficient explanation of deviations from predicted causal power levels.

References

Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104(2), 367-405.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic model of causal induction. Journal of Personality and Social Psychology, 58(4), 545-567.

Buehner, M. J., & 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.

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(6), 1119-1140.

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

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). Psychological Review, 107(1), 195-212.

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

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. Quarterly Journal Of Experimental Psychology Section B-Comparative And Physiological Psychology, 37(1), 1-21.

Vallée-Tourangeau, F., Murphy, R. A., Drew, S. (1997). Causal judgments that violate the predictions of the Power PC theory of causal induction. In M. G. Shafto and P. Langley (Eds.), Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society (pp. 775-780). Mahwah, NJ: Erlbaum.