uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Matthew Jacques
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Psychology)
GRADE / DEGREE

School of Psychology
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Covariation, Base-rate and Causal Power in Human Contingency Judgement

TITRE DE LA THÈSE / TITLE OF THESIS

Pierre Mercier
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Georges Fouriezos

Sylvain Gagnon (absent)

Rick Mehta

Catherine Plowright

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Covariation, Base-Rate and Causal Power In Human Contingency Judgment

Matthew A. Jacques

Dissertation submitted to the Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements for the PhD degree in Psychology

School of Psychology
Faculty of Social Sciences

Your file Votre référence
ISBN: 978-0-494-25876-7
Our file Notre référence
ISBN: 978-0-494-25876-7

To Anne, my reason and my heart.

Acknowledgements

This dissertation stands as the final product of my Ph.D. studies at the University of Ottawa. While I am very proud of this personal accomplishment, it would not have been achieved without the crucial guidance and collaboration of my supervisor, Dr. Pierre Mercier. In 2001, he welcomed me into his research lab and helped with my transition into what for me was a new area of research. By holding informal lab lunches on a regular basis, and by arranging an advanced seminar course in cognition, Dr. Mercier fostered an atmosphere which was both relaxed and intellectually stimulating. Now with a completed story to tell through my research, I cannot thank him enough for his role in guiding me through to this completed objective.

My committee has also played a critical role, and I thank Drs. George Fouriezos, Sylvain Gagnon, and Catherine Plowright for their timely responses, continual encouragement, and helpful feedback on earlier drafts of this document.

I have also been fortunate enough to have had several research assistants over the course of this project. Their help with coordinating the data collection has been invaluable. I would like to thank Katherine Brunner for her assistance with Experiment 2, Jennifer Quintal for Experiment 3, and Gary Capstick for his help with Experiment 4.

Finally, all of my family and friends have been remarkably understanding and patient with me throughout this entire process. Thank you so much – yes, it's finally done!

Abstract

This dissertation begins with a review of competing theories of human contingency judgment, and then describes a series of four original experiments designed to systematically investigate the strengths and weaknesses of current models. Among these, Cheng's (1997) Power PC theory of human contingency judgment has risen to prominence. This theory is said to address the inadequacies of both earlier associative (Rescorla & Wagner, 1972) and computational (Jenkins & Ward, 1965) approaches. Several prior tests of the Power PC theory have returned mixed or inconclusive results (Lober & Shanks, 2000; Vallée-Tourangeau, Murphy & Drew, 1997). The four experiments presented here were designed to assess predictions of the Power PC theory which had yet to be fully empirically tested, as well as to expand our current understanding of causal reasoning.

Experiment 1 results were consistent with the Power PC predictions in terms of the *pattern* of participants' judgments, but not with regards to the *levels* of those judgments. Through a replication and elaboration upon conditions from Experiment 1, Experiment 2 determined that the level of the Experiment 1 judgments was not due to a ceiling effect, as could have been argued by Power PC proponents. Experiment 3 served to investigate whether the concept of reliability could possibly explain the observed deviations from Power PC, and indeed demonstrated a significant interaction of power and reliability. Despite this finding, the conjunction of Power PC and reliability is shown to be problematic. Experiment 4 confirmed that the results from experiments 1 through 3 can be expected to hold true in negative contingency space as well.

Following the empirical results, further discussion of the roles of causal power, ΔP, and reliability raises new questions and reveals a number of viable options for future research in this area. Taken as a whole, the results from these four experiments do not support the original Power PC theory, but do provide insight toward alternatives involving both confidence and reliability, which will provide a more comprehensive account of human contingency judgment.

Table of Contents

Chapter 1: Introduction

How do people decide that the frequency with which two events co-occur reflects an underlying cause, and not merely a random coincidence? This report addresses the fundamental question of how individuals come to form conclusions about this type of relationship. Every day, we make decisions about our activities based on the presence or absence of external cues in our environments. These cues can be anything we perceive - whether a unique texture we feel, a familiar face we see, the smell of smoke, a sweet taste, or an alarming sound. Typically, the cue will have occurred prior to the effect. For example, someone may eat a particularly sour piece of food at dinner and then, recalling that they have previously experienced undesirable reactions to sour foods, attribute that as the cause of the upset stomach they experience later that night. Another simple example of this phenomenon arises when someone reasons, "there are many clouds in the sky, it looks like it is going to rain, so I'll bring an umbrella to work today". In this case and in myriad similar situations, the presence of one event (the cloud) leads us to predict the presence of another (rain). This type of relationship can be described as a binomial categorical contingency - the clouds are either present or absent and the rain is either present or absent.

When can we say that the relationship between two such cues is causal in nature, such that one would rationally and confidently conclude that *clouds cause rain* or *sour food causes stomach-aches*, rather than deciding that a purely coincidental or spurious correlation exists between the two? In other words, how do we decide which contingencies warrant our attention? After all, another observer might note that rain seems to coincide with or even follow after they see people wearing long-sleeved shirts!

Very few individuals would give credence to the idea that wearing long-sleeved shirts causes rain, although perhaps a similar belief is held by those few stalwarts who never fail to be seen wearing shorts when others have long-since reached for their winter clothes; perhaps they hold a belief such as "it will get cold and start snowing if I switch from wearing shorts to pants". This preposterous and marginally humorous example serves to highlight the distinction between correlation and causality. While there may always be curious exceptions such as the above, this research sets out to determine the fundamental rules and patterns which underlie normal human causal attribution.

As a most basic pre-requisite, relations can be said to be causal when there is an accepted or hypothesized mechanism which brings about the contingency. In the rain-clouds example, most observers have some level of understanding of the causal mechanism, whether on an overly simplistic level ("clouds are rain factories in the sky"), or a more sophisticated understanding ("the collision of a warm front with a cold front accelerates the condensation of airborne water droplets"). Non-causal contingencies are those that co-vary without any such mechanism to explain the covariation. In the absurd examples above, there is no known or even plausibly-hypothesized mechanism that I am aware of by which wearing long sleeves or pants can cause rain or snow. All contingencies can be considered either causal or non-causal depending on the presence or absence of such an underlying mechanism through which the occurrence of the first event can bring about the second. The focus of this report will be on the causal contingencies, which have a known or hypothesized mechanism of action by which the cause brings about the effect.

The wider topics to which the experiments presented here belong - causal judgment, associative learning, and decision-making - are entrenched in the histories of philosophy, psychology and cognitive science. From David Hume's 18th-century assertion that "causality is the cement of the universe" (Mackie, 1974), through to modern debates over artificial intelligence and machine learning, the principles under investigation here have both inspired and eluded researchers from many fields for centuries. Our ability to infer causality is so fundamental that it can be said to underpin even the most basic characteristics of what it means to be human – self-awareness, theory of mind, and intentionality. How do I know who I am, or that I even exist, without some understanding of my potential to purposefully interact with, and bring about effects in my environment? Among the historically significant descriptions of causal inference are found the ideas of Philosophers (Locke's *Essay on Human Understanding*, 1689; Hume's *Treatise on Human Nature*, 1740), Developmental Psychologists (Piaget's *The Child's Conception of Physical Causality,* 1930; Kohlberg's view of *The Child as a Moral Philosopher*, 1968), and in the writings of phenomenologists like Husserl and Heidegger, who reflect on how causal attribution plays a key role in how we as humans create meaning in the world around us (Iturrate, 1976; Nissim-Sabat, 1999).

Aristotle proposed four different types of causality: material cause, formal cause, efficient cause, and final cause (Horne, 2002). Material, formal and final causes from Aristotle's perspective generally deal with explanations of the static state or condition of objects, whereas efficient causes explain how changes are brought about. It is this efficient cause which is most similar to the causality discussed here; how something which occurs may be said to produce a change (i.e., clouds are an efficient cause of rain).

Hume (Mackie, 1974) would likewise propose that efficient causes are the most important ones, while effectively denying the existence of any objective, deterministic causality outside of the subjective perception of relations between events. Drawn to its ultimate conclusion, Humean causality implies that we cannot truly *know* that one thing causes another on the basis of past experiences, such that one cannot *know* that the sun will come up tomorrow. Instead, we must rely on the data that we acquire from our senses, and formulate our expectations about the future.

The philosophical tradition of causality was passed on and revised through early traditions in psychology, with no more prevalent example than Freudian and Jungian re-examining of the erstwhile Aristotelian 'efficient causes' behind neuroses and psychoses (Horne, Sowa and Isenman, 2000; Thoma and Cheshire, 1991). More recently, other researchers in cognitive psychology (Zorzi, 2004), developmental psychology (Phillips, 2004), cognitive science (Pacherie, 2004), and even artificial intelligence (Romdhane and Ayeb, 2005) have sought insights into these same processes of the mind, applying new and original methods of inquiry, but still possessing a desire to understand and predict the same phenomena. These pursuits emphasize how the phenomenon of attributing causality to people or objects is a fundamental process of human cognition present from the earliest stages of development, how it is considered essential for the abstraction of cognitive processes into developing models and theories of how the mind functions, and how any artificial mind or intelligence would be unconvincing without incorporating a realistic mechanism to simulate human contingency judgment. Considering the diverse intellectual fields with vested interest in the topic, the potential applications are far-

reaching for any validated theory which leads to improved predictions and understanding of causal attribution and decision-making.

The ontological and epistemological approach to be employed here emerges from the traditions of cognitive psychology and cognitive science. Two of the traditional core methods of inquiry in this domain include psychophysics and psychometrics. Psychophysics primarily concerns the desire to understand how individuals gather and process sensory information about their environment. This field was born of the initial attempts at quantifying and measuring human perception. Psychometrics represents the attempt to turn this process back on the individual, as researchers seek to understand and measure the otherwise unobservable internal, mental world, rather than the external environment. The experimental task described in this report will rely on both the presentation of external stimuli to the participants, as well as a self-reported value they generate after reflecting upon each trial, which represents the perceived causal strength present in the relation between the cause and effect over a series of trials. The causal strength reported by participants is a psychometric construct, as 'causality' does not physically exist outside of the mind of the beholder. Indeed, one person's perception of coincidence is another's idea of causality.

Recall the example of the individual who might believe that wearing long sleeves brings about cold weather, or perhaps yet another individual who believes that popsicle sales are somehow responsible for sunburns. Both of the presumed effects in these instances occur much more frequently when the presumed cause is present, yet very few of us would attribute a causal connection between the two. While these examples represent relatively extreme or obvious misinterpretations of contingency information, we

could all potentially hold any number of less outrageous yet still mistaken beliefs which arose from similar circumstances. By studying the process of causal attribution, we can come to more fully understand how both 'correct' and 'incorrect' subjective causal decisions are made. The experimental investigation of this psychometric construct of causality goes beyond the psycho*physical* question of whether and how the individual stimuli are received and encoded by the brain. The psychometric aspect of this type of research allows us to study higher cognitive processes such as how we create meaning from otherwise neutral elements in our environment, and how we generalize from a focal set of experiences to make decisions about the relations between people, places and things in our everyday surroundings.

A mathematical representation of contingency

To facilitate the study of causal attribution, it is possible to summarize contingencies as 2x2 tables, which represent a cross-tabulation of the presence or absence of the 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 a 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 the relation or contingency between the cause and the effect in this table can be quantitatively expressed as the ΔP coefficient (Equation 1). In this model, the contingency of the candidate cause and the effect 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. In this format, 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 $p(e|\sim i)$ 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 $p(e|\sim i)$ 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*).

$$\Delta P = \frac{a}{a+b} - \frac{c}{c+d} \tag{1}$$

$$\Delta P = P(e \mid i) - P(e \mid \sim i) \tag{2}$$

Following these formulations, it is possible to compute both positive and negative values of ΔP. A positive contingency would be calculated when the effect is more likely to occur in the presence of the cause. In such conditions where the hypothesized causal mechanism is intended to increase the likelihood of the effect, this can also be 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. If the effect occurs equally as often in the absence of the cause as it does in the presence of the cause, this would result in a ΔP value of zero, or a null contingency.

It should be noted that other models have been devised to investigate multiple candidate causes and effects, as well as causes and effects which are coded as continuous rather than binary events (Shanks, 1985a; Anderson and Sheu, 1995). While the insights gained from such approaches are tangentially relevant to the present research, the focus here will be on the unresolved questions relating to how humans attribute causality from fundamental, basic 2x2 binary-level contingencies. The processes involved at this basic level remain to be discovered, and the implications from such findings are expected to carry a high degree of generalizability to multiple-cue, continuous cause-effect relations.

Associative versus computational models of causal attribution

Most theoretical models of human contingency judgment fall into two major categories – computational or associative (Allan, 1993, Lober and Shanks, 2000). Purely associative models find their roots in Pavlovian classical conditioning. Such models describe a learning curve of increasing associative strength between a cause and effect,

on a trial-by-trial basis. This also relates back to Humean efficient causality which, as described earlier, assumes that humans cannot truly "know" that one event causes another on the basis of prior experience, and therefore must rely on external sensory information to formulate rational predictions about the future. 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. This model was originally intended to describe the learning process of animals, but it has since been widely applied to the study of human learning and causality judgment (Alloy and Tabachnik, 1984; Pearce and Bouton, 2001). Young (1995) emphasizes how there are many parallels between the causal learning processes of both animals and humans. Among the factors which influence both are temporal priority, temporal and spatial proximity, contingency, and prior experience (Young, 1995). While much of the early experimental knowledge of fundamental learning processes was based on animal studies, there is much to be gained by moving from analogy to direct human study. It has been a priority for many associative researchers to answer the question of whether the original utility of models such as Rescorla and Wagner's (1972) carries over to the human realm.

According to the associative perspective on human contingency judgment, participants do not compute the value of ΔP as described in equations 1 and 2, but rather are conditioned by the increasing association between cause and effect after repeated exposure to the four possible combinations - cause present or absent, with effect present or absent (Allan, 1993; Cheng, 1997; Lober and Shanks, 2000). As revealed in more detail in the following section, this type of model does well to explain the acquisition curve of the learning process, but has been faced with criticism for relying on seemingly

nebulous values for cause and effect salience (Cheng, 1997). The basic associative model on its own describes a dynamic process and was not intended to provide static, final predictions for causal strength. This has posed a challenge for arranging direct model-to-model comparisons, particularly to ΔP and computational models which by their nature provide *only* a final causal strength and do not describe the learning curve.

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.

The Rescorla-Wagner (1972) model

The Rescorla-Wagner (R-W) model (1972) has long been a leading candidate among associative models of contingency judgment (Allan, Tangen, Wood, and Shah, 2003; Cramer, Weiss, William, Reid, Nieri, and Manning-Ryan, 2002; Lober and Shanks,

2000; Yamaguchi, 1999; Shanks, 1987). The original conceptualization dealt with the association between a conditioned stimulus and an unconditioned stimulus in traditional Pavlovian animal conditioning tasks. However, the goal of this model when applied to the human population is to describe the learning curve as a participant is exposed to repeated pairings of a candidate cause and its presumed effect. Here the candidate cause takes the place of the conditioned stimulus, while the presumed effect takes the place of the unconditioned stimulus. The R-W account of the associative learning process has been used to explain several phenomena of human causal learning. Traditional blocking, overshadowing and inhibition effects[1] are all explainable by the standard R-W formula (Shanks, 1985b; Lober and Shanks, 2000).

In its typical formulation, Equation 3 describes the change in associative strength between a candidate cause and effect on any given trial. Thus, the components of the R-W model are: ΔV_c, which represents the change in associative strength of a given cue on a given trial; α and β are salience weightings for the candidate cause and the effect respectively; λ represents the asymptotic or maximum possible associative strength, assumed to be 1 in the presence of the effect and 0 in its absence; and ΣV represents the sum of the associative strengths prior to that trial. The specific weightings for α and β can vary depending on whether the cause and effect are present or absent. The R-W model has been criticized for providing no firm guidelines for setting these salience weightings, however they must fall somewhere between 0 and 1.

[1] These three effects all refer to means by which the conditioning or learning of the cause-effect relationship is undermined. Blocking refers to cases where another cause is already strongly associated with the effect. Overshadowing occurs when a more salient stimulus occurs simultaneously with the candidate cause. Lastly, inhibition takes place when there has been significant pre-exposure to the candidate cause in the absence of the effect, thereby reducing the potential for conditioning when the cause is paired with the effect.

$$\Delta V_c = \alpha_c \beta_e (\lambda - \Sigma V) \tag{3}$$

By extrapolating this formulation for the change in associative strength on a single trial to an entire series of trials, a learning curve can be generated. This curve progresses as a series of error-correcting steps, from a starting point of zero towards a value which represents the sum of all of the ΔV_c values for the total number of trials. This curve is determined by Equation 4, below. The associative strength present on the next trial is equal to the associative strength on the current trial plus any change brought about by the current trial (as defined by Equation 3). In theory it is possible for the associative strength to fluctuate indefinitely over an infinite series of trials, however as the number of trials increases, the amount of variance produced by a single trial becomes negligible.

$$V_{c(n+1)} = V_{c(n)} + \Delta V_c \tag{4}$$

In the R-W model, this asymptote is reached as causal reasoners have processed a sufficient number of trials to have perceived the underlying relation between the candidate cause and effect. Let us return to one of the earlier examples of someone experiencing an upset stomach many of times after eating a particular type of food. Their belief in the causal strength of that food to cause their stomach-ache is unlikely to waver over the next one or two times they eat the food, no matter the result. At that point it could be said that their judgment of the causal strength has reached asymptote. The numerical value that this individual would place on that causal strength may be of significance to investigators who are striving to understand the nature of the causal process in research such as will be presented here, but the individual for their own part might simply experience the causal connection as a belief like "I'll never eat that food again", or "It tastes delicious, so it's worth getting an upset stomach".

The fact that the R-W model describes the trial-by-trial changes in association, and predicts the subsequent learning curve, also creates some potential limitations (Lober and Shanks, 2000). The ΔP and PCM approaches are computational, and represent attempts to answer the question "*what* do people calculate when asked to make a causal judgment?". On the other hand, the R-W model is akin to an algorithm which describes the process, "*how* do people make decisions about causality?". At one level, these questions are fundamentally different from each other, yet several researchers have found a common ground which has allowed for the differing approaches to be compared empirically. One logical step has been to use the final, asymptotic predictions of the R-W model to use for comparison to computational models. This allows the predicted pattern of both to be compared to each other. Interestingly, Cheng (1997) has also revealed that the asymptotic predictions of the R-W model are equal to those of the PCM, as long as the β weightings for conditions where the effect is present and absent are equal. As such, when these requirements are met, the PCM's ΔP metric can be used as a proxy of the R-W model's asymptotic judgments.

However, the equal-β prerequisite of Cheng (1997) has been met with some contention, as the original R-W model (Rescorla & Wagner, 1972) makes no claim that the β weights must be equal. Lober and Shanks (2000) thus refer to the equal-β version of the R-W model as the *restricted* R-W model, and any variation which allows the β weights to be unequal as *unrestricted* R-W. Given this distinction, while the R-W model may produce predictions which are consistent with ΔP in some circumstances (restricted model), this should not be taken to mean that the R-W model as a whole necessarily reduces to ΔP for all possible values of effect salience (β).

Cheng's (1997) Power PC theory

While there is a history of empirical support for most of the predictions of ΔP and the PCM (Cheng and Holyoak, 1995; Shanks, 1985b; Shanks and Lopez, 1996), 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 would be calculated as 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 (Wu and Cheng, 1999; Cheng, 1997).

To illustrate these situations, Cheng described two scenarios. In the first, a doctor would like to know if a patient is allergic to certain substances. To test for allergies, the doctor lightly scratches the patient's skin a number of times and then applies the various allergens to the scratched skin to test for a reaction. If all of the allergens produce a swelling reaction ($p(e|i) = 1$), yet this particular patient's skin just happened to be so sensitive that the initial scratching would have been sufficient to produce the same swelling ($p(e|\sim i) = 1$ also), the doctor would fail to make any judgment about the allergens if they were basing their reasoning upon the PCM, as ΔP would be zero. The second scenario involves a researcher who would like to determine the effectiveness of a medication to relieve headaches. One group receives the medication, while a second receives a placebo. If no differences are found between the two groups, the researcher

might conclude that the medication was ineffective if they were reasoning on the basis of the PCM. However, it might be possible that neither group experienced any headaches (when $p(e|i)$ and $p(e|\sim i)$ both = 0). The two examples reveal how for generative causes in which the effect always occurs (first example), or for preventive causes in which the effect never occurs (second example), the PCM could mistakenly conclude that the candidate cause fails to bring about the effect.

A number of other researchers (Allan and Jenkins, 1983; Baker, Berbrier and Vallée-Tourangeau, 1989; Buehner and Cheng, 1997; Shanks, 1987) have revealed how participant judgments can deviate from the ΔP metric which forms the basis of the PCM. These studies found that an increasing overall probability of the effect, $p(e)$, will lead participants to evaluate a candidate cause to be significantly higher than candidate causes with a lower probability of the effect. This finding was strongest where $\Delta P = 0$, and an increase in probability of the effect consistently resulted in higher causal judgments.

To address these 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 judgment. The concept of causal power was introduced as an improvement over purely co-variational models such as ΔP and the PCM, as well as the associative account as represented by the Rescorla-Wagner model. The Power PC theory itself actually incorporates elements of the original PCM (namely, ΔP), and indeed the 'PC' still refers to *probabilistic contrast*. Buehner, Cheng and Clifford (2003) argue that many of the weakness of earlier models (including the R-W model) resulted from failing to heed the admonition that correlation does not equal causality. In their view, the PCM and R-W are

limited in that they hold covariation to be a direct measure of causation. The Power PC theory, on the other hand, was designed clarify the distinction between the two and provide a testable metric for human *causal* induction. 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). This implies that there must be some known or hypothesized causal mechanism by which the cause brings about the effect, and that the goal of the causal reasoning task is to inductively form a conclusion about the strength of this mechanism on the basis of the evidence available to them. The computational equation of the Power PC theory for generative causes is:

$$p_i = \frac{\Delta P}{1 - p(e|\sim i)} \tag{4}$$

In Equation 4, ΔP represents the level of contingency between the cause and the effect (as in the PCM), $P(e|\sim i)$ 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 (a) of the effect (Cheng, 1997):

> "(1) When i occurs, it produces e with probability p_i; when a occurs, it produces e with probability p_a; and nothing else influences the occurrence of e;
>
> (2) i and a influence the occurrence of e independently; and
>
> (3) i and a influence the occurrence of e with causal powers that are independent of how often i and a occur"
> (Cheng, 1997, 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:

$$p_i = \frac{-\Delta P}{p(e\,|{\sim}\,i)} \tag{5}$$

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 4) and negative (preventive, using Equation 5) contingencies. For any given value of ΔP (left side of table), with a specified base rate of $P(e|{\sim}i)$ (along the top of table), the predicted causal power levels are provided.

Table 2: Contingency and Causal Power Matrix

	P(e|~i)										
ΔP	**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	0.00*
-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

*Note that causal power is only determinable in this condition if the preventive formulation is used.

Cheng (1997) has described how, despite serving as an improvement to the PCM, in many circumstances, the Power PC theory reduces to ΔP. This can be seen in generative contingencies where $p(e|{\sim}i) = 0$, rendering the denominator of Equation 4 equal to 1, and p to ΔP, as well as preventive contingencies where $p(e|{\sim}i) = 1$, which reduces Equation 5 to ΔP. In these conditions, several of the phenomena explained by the PCM, such as blocking, overshadowing, and conditioned inhibition, are equally explainable by the Power PC theory (Shanks, 1985b; Cheng, 1997). Lober and Shanks (2000) also highlighted how Cheng's (1997) derivation of the Rescorla-Wagner model which requires equal salience weightings is equivalent to the PCM. This means that for conditions where the Power PC model reduces to ΔP, the restricted R-W also makes identical predictions.

Despite this interesting consideration of equal predictions, the focus of this present study will be on conditions where the models are more readily differentiated. As shown in the Table 2, at somewhat extreme values of low contingency ($\Delta P = .2$) and high base rate ($P(e|{\sim}i) = .8$), the predictions of the Power PC theory, namely that causal power (p) tends toward 1.0, appear implausible. Intuitively, it seems unlikely that participants would judge conditions of low contingency as having maximal causal power. Why would an event that is <u>not</u> systematically followed by a certain effect be judged to strongly cause that effect, even or especially when the effect occurs frequently on its own (i.e., there is a high base rate of the effect)? There is some previous research which has examined p_i as a function of base rate. For instance, Buehner and Cheng (1997), Vallée-Tourangeau *et al.* (1997, 1998), and Lober and Shanks (2000) have indeed shown a relation between $P(e|{\sim}i)$ and perceived causal strength. However, they have not comprehensively explored

the many predictions contained in Table 2 and the results have been conflicting. 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 types of situations, it is puzzling that Power PC predicts maximal contingency judgments, yet if empirical support can be generated for those predictions, the model would be established as a significant improvement upon earlier associative and computational approaches. The next section provides more detail of the rationale and results of the earlier studies which set out to either support or discredit the Power PC theory of human contingency judgment.

Prior Tests of the Power PC theory

There have been several early tests of the Power PC theory of human contingency judgment since it was introduced (Buehner and Cheng, 1997; Lopez, Shanks, Almarez and Fernandez, 1998; Vallée-Tourangeau, Murphy and Drew, 1997) Most of these contrasted the predictions of the Power PC theory against those of the PCM and R-W models by varying $p(e|i)$ and/or $p(e|\sim i)$, while holding ΔP constant across a series of trials. The other logical approach to this task is to hold p_i constant while varying the levels of ΔP in each experimental condition (Vallée-Tourangeau, Murphy and Drew, 1997). The findings of several key experiments often cited in the debate over the Power PC theory will be summarized in this section.

Buehner and Cheng (1997) designed a human contingency judgment experiment which combined five levels of $p(e|i)$ with five levels of $p(e|\sim i)$, for both generative and

preventive causes. The levels of $p(e|i)$ and $p(e|\sim i)$ were .00, .25, .50, .75, and 1.0, which in combination also produced ΔP values of .00, .25, .50, .75, and 1.0 for generative causes and 00, -.25, -.50, -.75, and -1.0 for preventive causes. In the first experiment examining the preventive power of vaccines, Buehner and Cheng were interested in determining whether participants' judgments would decrease over a constant ΔP with increasing base rate $p(e|\sim i)$. The experimental task for this experiment asked participants to pretend they were virologists testing a series of vaccines to prevent diseases. The second experiment concerned whether or not generative conditions with constant ΔP would show an increase as base rate $p(e|\sim i)$ increased. Here participants were asked to take on the role of a microbiologist studying the role of ray exposure on the development of viruses. For each condition, participants were asked to predict on a trial-by-trial basis whether the effect would be present on the basis of a fictional laboratory record which indicated the presence or absence of the supposed cause. Final judgments were also collected after each block of sixteen trials. These ratings were gathered on a scale from 0 to 100, with 0 indicating that the cause is completely ineffective in bringing about the effect (whether generative or preventive), and 100 indicating that the cause never fails to bring about the effect.

Overall, the participants' judgments for both generative and preventive causes varied systematically with $p(e|\sim i)$ within any given value of ΔP. This finding is consistent with the Power PC theory in that causal power is fundamentally a systematic deviation from ΔP on the basis of $p(e|\sim i)$ (see Equations 4 and 5, above). However, participant judgments also deviated from 0 for all cases where $\Delta P = 0$. This finding is consistent with a known weakness of the PCM discussed earlier (Allan and Jenkins, 1983; Shanks,

1987), where increasing overall probability of the effect seems to moderate judgments away from zero; however it is a result which remains unexplained by any of the PCM, R-W or the Power PC theory on their own. Buehner and Cheng (1997) argued that these deviations were due to the number of trials being limited to 16 in both experiments. They report two follow-up experiments where data for 100 trials were presented to participants in summary form, and the judgments correspondingly returned to a lower level consistent with the Power PC theory. Unfortunately, the presentation of data in the additional experiments was done in summary form, a change in experimental modality which introduces a confound. On the other hand, the deviations from zero in the first two experiments remain as evidence that perhaps the participants' judgments had indeed not reached asymptote after 16 trials.

The systematic deviations from ΔP reported by Buehner and Cheng (1997) provide evidence against the basic PCM. They also claim that the results are inconsistent with the R-W model. They indicate that a restricted (equal salience values for effect present and absent) R-W model cannot account for the pattern of results, and that only applying different salience weights for generative and preventive conditions will allow the R-W model to account for the results. As there is no single set of R-W salience parameters which fits the data, Buehner and Cheng (1997) conclude that even an unrestricted R-W model is ruled out. However, Rescorla and Wagner (1972) never argued that salience weights must be equal, or that only one set of salience weights must be applied. Indeed, the Power PC theory itself utilizes two separate formulas (Equations 3 and 4) to account for the causal attribution process in generative and preventive scenarios.

Vallée-Tourangeau, Murphy and Drew (1997), took a different approach than Buehner and Cheng (1997). Rather than holding ΔP constant and varying $p(e|i)$ and $p(e|\sim i)$ (effectively varying the Power PC's p_i), they chose to isolate p_i while varying the values of ΔP. Thus, the experimental set up called for conditions where the Power PC theory would predict constant levels of causal power, while the PCM would expect participants' judgments to vary. The participants were asked to assume the role of a researcher who wanted to determine whether certain hormones either caused or prevented cancer in mice. Both positive and negative values of ΔP were used in the experiment, with participants providing a rating for each trial on the basis of a scale ranging from -100 (hormones prevent cancer) to 100 (hormones cause cancer). Two experiments were carried out: one with p_i set to 1.0 or -1.0, with ΔP values of +/- .83 and .17, and another with p_i set to .50 or -.50, with ΔP values of +/- .50 and .10. In the first set, both generative and preventive scenarios resulted in participant judgments which deviated from the predicted p_i values of 1.0 and -1.0. In addition, the preventive conditions mirrored the generative conditions almost perfectly, with mean ratings of +85.9, +61.8 for the generative conditions of ΔP = .83 and .17 respectively, and mirrored by -85.4 and -61.5 for the preventive conditions of ΔP = -.83 and -.17.

In the second experiment from Vallée-Tourangeau, Murphy and Drew (1997), where power was set to +/- .50, a surprising result was reported. Similar to the first experiment, the participants' judgments deviated significantly from the predictions of the Power PC theory. In addition, the judgments departed from the predictions of the PCM and R-W models. For conditions of ΔP = +/- .50, participants' ratings were 24.4 and -10.2 respectively. For conditions of ΔP = +/- .10, ratings were 43.8 and -49.8. There is

currently no model or theory of human contingency judgment which would adequately account for the results of both experiments. Particularly challenging is the result that small values of ΔP returned more extreme causal strength ratings than those conditions where ΔP was moderate.

Lober and Shanks (2000) sought to clarify some of the confounding results from both Buehner and Cheng (1997) and Vallée-Tourangeau, Murphy and Drew (1997). Three experiments were conducted which utilized a similar experimental protocol to both of the earlier studies. The particular cover story in this instance was for participants to pretend they were researchers working in a laboratory investigating the effects of certain chemicals on animals' DNA. They employed a strategy whereby p_i was held constant for each of the first two experiments, but varied in the third. Ratings were collected on a similar scale of 0 to 100, where 0 indicates that the chemical does not cause a DNA mutation, and 100 indicates that the chemical always causes a mutation. The first experiment had constant $p_i = .70$, with ΔP values of .70, .47, and .23. Participant judgments were found to vary significantly away from p_i, in direct relation to the ΔP values. This contradicts the Power PC theory, and supports both the PCM and R-W.

Lober and Shanks' (2000) second experiment was designed with a more extreme value of p_i set to 1.0, while ΔP levels were 1.0, .75, .50, and .25. Again, mean judgments were at direct variance to the Power PC predictions, and significantly demonstrated a linear relation to ΔP. For this experiment, Lober and Shanks also noticed a sub-set of participants who seemed to be following the Power PC theory rather than the PCM. A post-hoc analysis, separating the participants into Power PC and PCM adherents, revealed a significant difference between the two groups. This finding seems to indicate

that individual participants may arrive at their contingency judgments via different strategies.

The third experiment from Lober and Shanks (2000) attempted to hold ΔP constant while varying the predictions of the Power PC theory. Contingency levels of ΔP = .40 was used for three of the four trials in this experiment, with the fourth using .07. Causal power levels of p_i = 1.0, .67, .40 and .40 were used. In direct contrast to their first two experiments, the results here indicated that participants were formulating their contingency judgments on the basis of p_i rather than ΔP. There was a statistically significant trend of p_i, while no analyses resulted in a main effect or interaction with ΔP. The experimenters were somewhat at a loss to explain the results of this experiment in light of the first two. They suggested that some combination of salience weights from an unrestricted R-W model might be able to fit the results. Likewise, applying different weighting parameters to the $p(e|i)$ and $p(e|\sim i)$ components of the PCM also allows for the results of all three experiments to be accounted for. However, Cheng (1997) and Buehner, Cheng and Clifford (2003) have denounced the alternate-weightings strategy as being a post-hoc attempt to fit the data at hand, while their Power PC p_i value is intended to be normative and not require any such adjustments.

Shanks (2002) later set out to specifically test the predictions of the Power PC theory in negative contingency space. While others (Vallée-Tourangeau, Murphy and Drew, 1997) had examined a few conditions of preventive causal power, there had yet to be a deliberate investigation of participants' responses to preventive scenarios while holding p_i constant. Again, participants were asked to assume the role of a researcher who wants to determine whether certain chemicals have an effect on animals' DNA. The

first experiment included two conditions which fixed p_i at -1.0, with ΔP = -.2 and -1.0. The second fixed p_i at -.7, with ΔP = -.23 and -.7. Both experiments also included a condition of p_i = .75, ΔP = .6 so that participants would have an opportunity to distinguish between generative and preventive conditions. For both experiments, judgments varied with ΔP in direct contrast to the Power PC theory. The second experiment also involved the collection of participants' confidence ratings, something which is still relatively rare in this type of research. Participants were slightly but significantly more confident in the ΔP = -.7 condition than they were in the ΔP = -.23 condition. This finding supports the argument raised by Buehner and Cheng (1997) and Buehner, Cheng and Clifford (2003), that participants may be modifying their judgments on the basis of how confident they are in making those judgments. However, Shanks (2002) dismisses this argument by revealing that the confidence difference between -.7 and -.23 was very small, and not enough to account for the otherwise large difference in causal judgments between the two conditions (which is NOT predicted by the Power PC theory).

The results of these first investigations of the Power PC's predictions are ultimately inconclusive. Some of the experiments seem to provide strong support for the Power PC theory, while others are either unclear or provide equally strong evidence for the PCM or R-W models. Because previous researchers have not fully and systematically explored the most counter-intuitive predictions of the Power PC theory, still 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 situations, it appears puzzling that Power PC predicts maximal

contingency judgments. If the Power PC theory is to be welcomed as a true normative standard of human contingency judgment, even its most counter-intuitive predictions must hold true. This approach to studying causal attribution appears to be ideal, in that it will either provide the supporting evidence currently lacking for the Power PC model, or it will suggest that the theory ought to be abandoned, while providing potential insights towards alternatives.

For contingency judgments to be maximal, as the Power PC theory predicts in these most counter-intuitive conditions, it is not necessary for them to reach an absolute numerical value of 100 or 1.0. Rather, they need only be demonstrated to have reached a subjective point that the participant would rate consistently as the highest. If no experimental conditions generate a higher contingency rating from the research participants, then it can be reasonably assumed that a maximum ceiling was reached. On the other hand, if one experimental condition generates causal attribution ratings which are statistically significantly higher than another condition, one could reasonably conclude that the lower of the two conditions was not in fact at a true maximum or ceiling.

Relation with confidence and reliability

This report focuses on a series of empirical tests of the Power PC's most counter-intuitive predictions in situations such as those described above. Participants will provide causality judgments in a computerized task similar to Cheng (2000). In parallel with causality judgements, the experiments will monitor confidence. 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; however 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 (Shanks, 1985a; Vallée-Tourangeau et al, 1997, 1998; 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 such as the PCM and Power PC provide an end-point for causal judgments, compared to the learning-curve described by associative accounts. Systematically 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. Since confidence has rarely been studied in this context before, it may prove to be a valuable addition or adjunct to one of the prevailing theories, or it may provide insights towards an altogether new account of the causal attribution process.

The relationship between confidence and contingency judgment has only begun to be studied. Clément, Mercier and Pastò (2002) have demonstrated how, counter-intuitively, participants' level of confidence in their judgments decreased as the number of trials increased, and varied in direct proportion to the normative ΔP values. This would seem to be at variance with the concept of an acquisition curve, as well as the associative account which describes judgments as reaching asymptote after a relatively large number of trials. If judgments are stabilized after a large number of trials, yet participants are increasingly unsure of their responses, are there new insights from this discovery which

ought to be incorporated into any successful account of human causal attribution? Shanks (2002) also collected confidence ratings, but dismissed a significant effect of confidence in favour of stronger evidence for a ΔP factor. This resulted in his rejection of the Power PC theory, but ultimately proves to be a weak investigation of the potential role that confidence may play in human contingency judgments.

Clément, Mercier and Pasto (2002) also describe how the relative simplicity or complexity of the distribution of *a-b-c-d* cells in the 2x2 contingency table may affect confidence. For a series of 40 trials, if all 40 fall within cell *a* (cause present | effect present), the contingency could be described as very simple. Conversely, if 40 trials are split up as 10-*a,* 10-*b,* 10-*c,* and 10-*d*, then the 2x2 contingency would be described as relatively complex. The research described here will build upon this information about how confident participants are in the contingency judgments they provide. It is hoped that other researchers will follow, and continue to collect data regarding participant confidence and the role it may play in causal attribution.

It is also in relation to confidence and its potential impact on expressed judgement that Buehner and colleagues (Buehner & Cheng, 1997; Buehner, Cheng & Clifford, 2003), have raised the issue of how the results of several contingency judgment experiments which were inconsistent with the Power PC theory could be due to participants confounding causal power with the concept of reliability. They explain reliability as the number of opportunities the candidate cause (*i*) has to demonstrate its effect (*e*), minus the number of trials in which the participant might expect the effect to happen simply by chance considering the base rate of the effect ($P(e|\sim i)$). In other words, when an effect occurs with increasing frequency in the absence of the candidate cause,

observers will need to observe more and more instances of the effect taking place in order to draw a conclusion about the causal power of the candidate cause. This represents a condition of low reliability. Conversely, when the effect occurs rarely or never in the absence of the candidate cause, individuals will require less exposure to the effect to draw their conclusion about the causal power of the candidate cause. This would be considered a condition of high reliability. Consider this example: it may be harder to determine if someone has pinched you while you are outdoors and being bitten by mosquitoes constantly (high occurrence of the effect in the absence of the cause) than if you are indoors (low occurrence of the effect in the absence of the cause).

While the exact formulation of reliability is not given in these sources (1997, 2003), M. Buehner (personal communication, October 5, 2005) has supported the formalizations of the construct I have derived as Equations 6 and 7 below. Equation 7 is simply an alternate notation, reflecting the consistent use of the a, b, c, and d terms.

$$R = (a+b)-(a+b)*P(e|\sim i) \tag{6}$$

$$R = (a+b)-(a+b)\left(\frac{c}{c+d}\right) \tag{7}$$

Equations 6 and 7 depict 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. Reliability is then intended to act as an adjustment or moderating factor within the Power PC theory, and currently exists as a post-hoc construct to account for experimental data which are inconsistent with Power PC alone. I have derived Equations 8 and 9 as a means of expressing reliability as a proportion rather than a raw number, by dividing the

number of trials from Equations 6 and 7 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 and compared across different experimental conditions. Again, Equation 9 is simply a restatement of Equation 8, arrived at by reducing $p(e|\sim i)$ to its most basic terms based on 2x2 contingency cell frequencies.

$$R = [(a+b)-(a+b)*P(e|\sim i)]/(a+b) \tag{8}$$

$$R = \left(\frac{(a+b)-(a+b)\left(\frac{c}{c+d}\right)}{(a+b)} \right) \tag{9}$$

It is hoped that our incorporation of reliability into the investigation of the Power PC theory will augment the current level of understanding about the role of reliability, on both judgment and confidence. If participants are confounding reliability with causal power, one would expect their confidence ratings to drop when reliability is low. The full implications of the use of reliability as an adjunct to the Power PC theory, and other considerations will be examined in more detail in the general discussion following the experimental results. As it stands, the formalization of reliability as a proportional construct in Equation 9 represents an original contribution to the research area, which Buehner (personal communication, March 3, 2006) has described has a welcome and necessary step beyond the descriptive level as provided by Buehner and Cheng (1997) and Buehner, Cheng, and Clifford (2003).

Chapter 2: Contingency Judgment Experiments

Rationale

To begin our investigation of contingency judgments, we ask "How would we know if research participants' perceptions of causality are consistent with, or inconsistent with, a particular model's predictions?" In a research environment, participant judgments would be considered consistent with a particular model if conditions with a high predicted ΔP or p_i are rated higher than judgments for medium ΔP or p_i conditions, which in turn are rated higher than low ΔP or p_i conditions. The results will be examined primarily on the basis of whether the *ordinal* pattern of obtained judgments is consistent with the predicted pattern. This also carries with it the implication that no direct empirical comparisons should be made between the *normative* or calculated values of ΔP and the *subjective* or reported values obtained from participants.

Making such comparisons would violate statistical assumptions of a common scale of measurement. Following from Stevens's (1946) seminal work on levels of measurement, we cannot assume that subjectively-generated ratings from participants adhere to the same measurement scale as the predicted, normative values. While we know that ΔP values of .9 or 1.0 represent a high level of cause-effect contingency, how can we know whether participants' internal perceptions and representations of that contingency follow the same rules? Or to pose a question paraphrased from a seminal example of mathematical scaling issues (Mandelbrot, 1967), how does one go about answering the question "How long is Canada's coastline?" There is a standard answer to this question, presented by Natural Resources Canada (2000), which can serve as a proxy to the computed, *normative* values of ΔP or Power PC. The coast of Canada is officially

202,080 Km (akin to the normative value of ΔP), however any individual's attempt at answering this same question will yield a different answer depending on which instruments they have at their disposal (i.e., measure every kilometre, or centimetre?). If measurements of one kilometre are used as the base unit, there will be an inherently larger 'margin of error', than if centimetre or millimetre units are employed. How far into every inlet or river on the coast will we measure? Though individual coastal measurements can be derived in myriad ways, and are likely to vary from the standard, normative answer, it is inappropriate to label them as "wrong". In this same manner, participant's judgments of contingency will not be analysed for "accuracy", rather their judgments on one set of experimental conditions will be compared to their own judgments on a different set of conditions. To refer back to the metaphor, individuals' measurements of the coastline of Canada would be compared to their own measurements of different countries' coastlines, while the researcher asks "Did the participants notice a significant difference between the coastlines?" and "What features of the coastlines led to their perceiving them as different?"

As a more direct illustration of this issue, we cannot know that the difference between a reported .05 and .10 contingency is as significant as the difference between .90 and .95. Whereas the normative values between 0 and 1 follow a linear scale with equal values for increments of .05 everywhere along the scale, the subjective reports may reflect a non-linear scale with different 'anchor points' – the difference of .05 between .05 and .10 might be completely arbitrary and meaningless if both are considered 'Very Low' to the participants, whereas the .05 between .90 and .95 could be reflect a categorical jump from 'Very High' to 'Extremely High'. Or alternately, the participants'

subjective scale may still be linear, yet it might have different absolute minimal and maximal values – if a participant reports a causal judgment of .80, how are we to know if that is meaningfully different from a predicted value of 1.0 (maximum)? Is the difference attributable to error, or is the participant merely stating that .80 is the maximum causal power they are willing to attribute to the candidate cause? It is for this reason that we are less concerned with the *accuracy* of contingency judgments per se than what the *pattern* of judgments can teach us. The questions to be examined by this research fall along the lines of "Which model's characteristics (PCM, Power PC) are most consistent with the participants' reported contingency judgments?" and not "Were the participants' judgments correct?"

To answer these questions, a repeated-measures factorial design was used in all of the experiments. Because all of the independent variables of interest (ΔP, p_i, $P(e|i)$, $P(e|\sim i)$, and R) were controlled for by manipulating the values of the a, b, c, and d cells, each participant could experience all of the experimental conditions and serve as their own comparison group.

Experiment 1

The purpose of Experiment 1 is to examine the patterns of causal judgments and confidence in situations of where the Power PC theory predicts either high or low power even when contingency remains low. The experimental conditions chosen for this first experiment were drawn from the more counter-intuitive cells in Table 2.

Method

Participants

Twenty-four volunteer undergraduate students (10 Male, 14 Female) from the University of Ottawa took part in the experiment 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. 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 icon measured approximately 9mm high and 5mm wide, while the face icon had a diameter of approximately 12mm. 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. Each of the four possible combinations are represented in Figures 1 through 4 below. The treatment (pill) and outcome (skin condition) were presented simultaneously, for the duration of 1 second, with an inter-stimulus interval of 1 second. The two pill colours – blue and red – were counterbalanced according to odd and even numbered experimental conditions. The

colours were only used to increase the generality of the results but were not included in the analysis as there was no theoretical hypothesis attached to them.

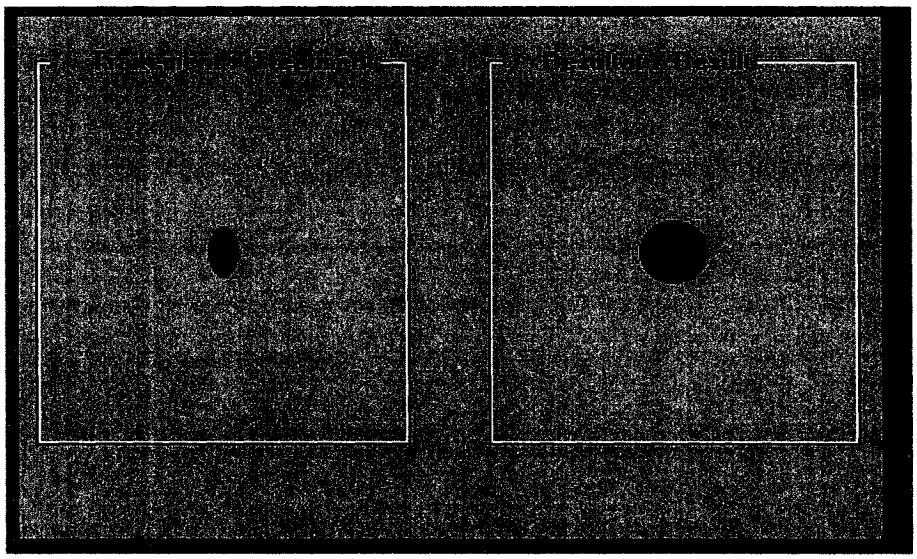

Figure 1: Medication Present and Skin Rash

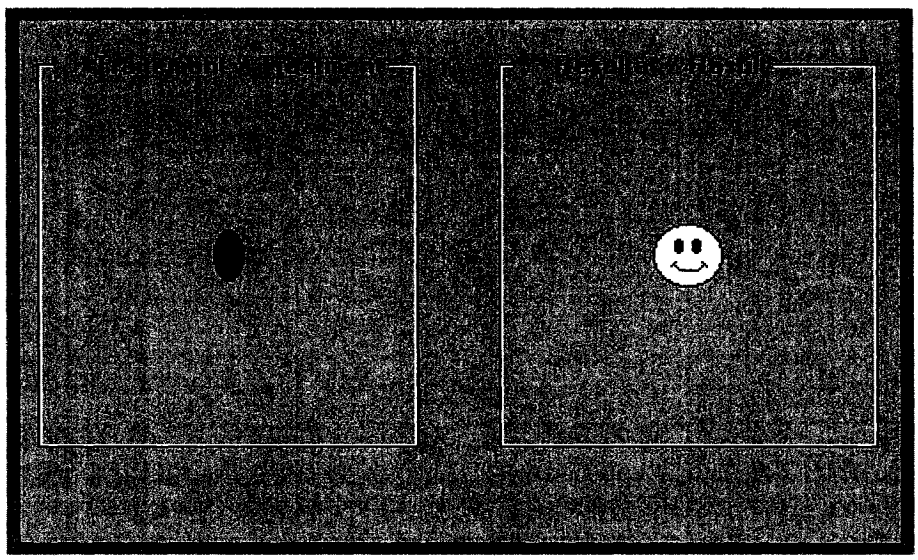

Figure 2: Medication Present and Happy Face

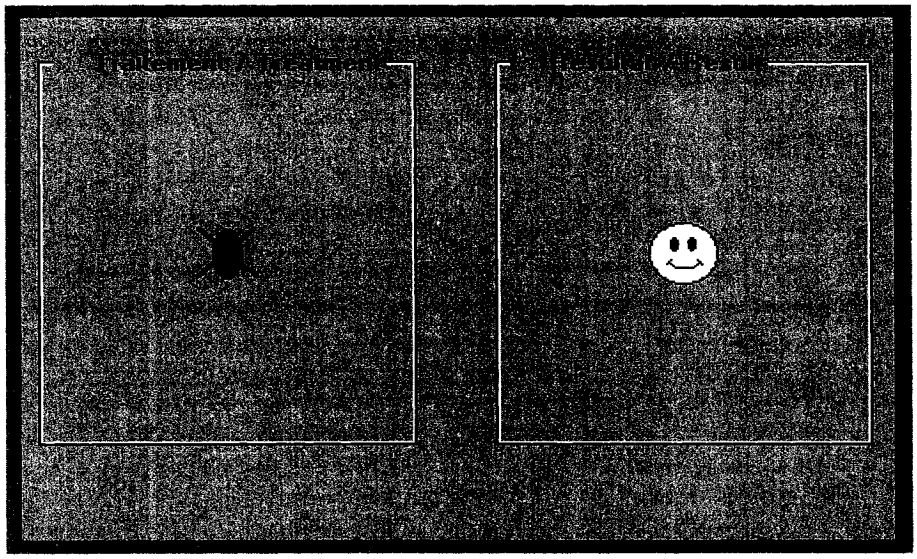

Figure 3: Medication Absent and Happy Face

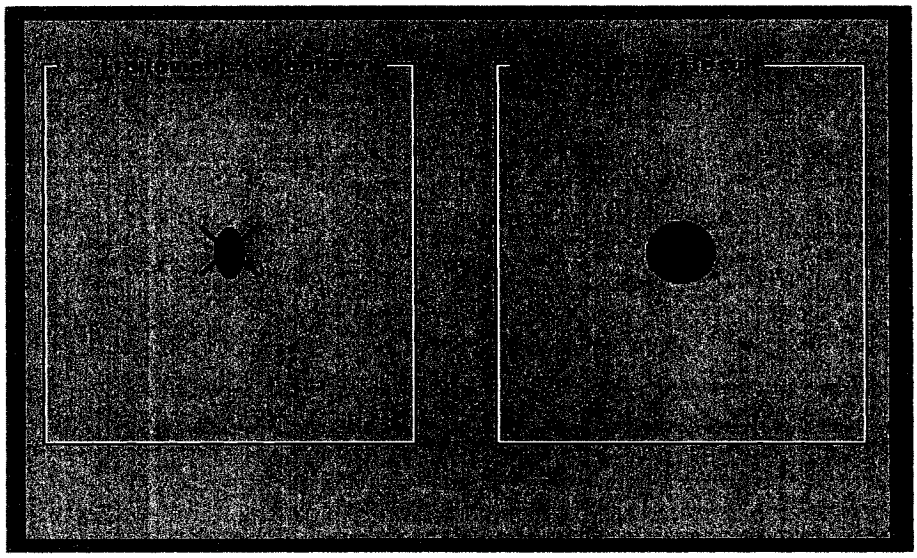

Figure 4: Medication Absent and Skin Rash

Procedure

Participants were seated in front of the computer and oriented to the experiment with written instructions which appeared on the computer's screen. At the same time, icons representing each of the possible conditions (presence/absence of medication, and

presence/absence of skin condition) were shown to familiarize the participants with the stimuli. 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."

The three practice conditions then followed, consisting of 40 medical files each. As listed in Table 3, the first practice had a ΔP of .67, the second -.67 and the third was 0. The order of the practice conditions was fixed and chosen to emphasize contrast in covariation and in power (.67, 1, and 0) simultaneously. Each condition included the presentation of a set of 40 fictitious medical files, each of which constituted a trial. Each

set of 40 trials was made up of two blocks of 20 trials keeping covariation and other variables constant. The order of presentation of the trials was randomized within each block of each condition. At the end of each condition, participants were asked to judge the relative efficacy of the treatment on a scale from -100 ("The risk of skin discoloration has decreased") to 100 ("The risk of skin discoloration has increased") with 0 labelled as "The risk of skin discoloration has remained unchanged". 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". This contingency judgment and confidence rating screen is represented in Figure 5, below. Participants were not given feedback regarding the accuracy of their judgments at any point in order to genuinely capture their subjective impression rather than shape their judgement. Nevertheless, the normative contingencies used in the practice conditions were chosen to expose each participant to a strong positive, a strong negative, and a null contingency before proceeding to the experimental trials. 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.

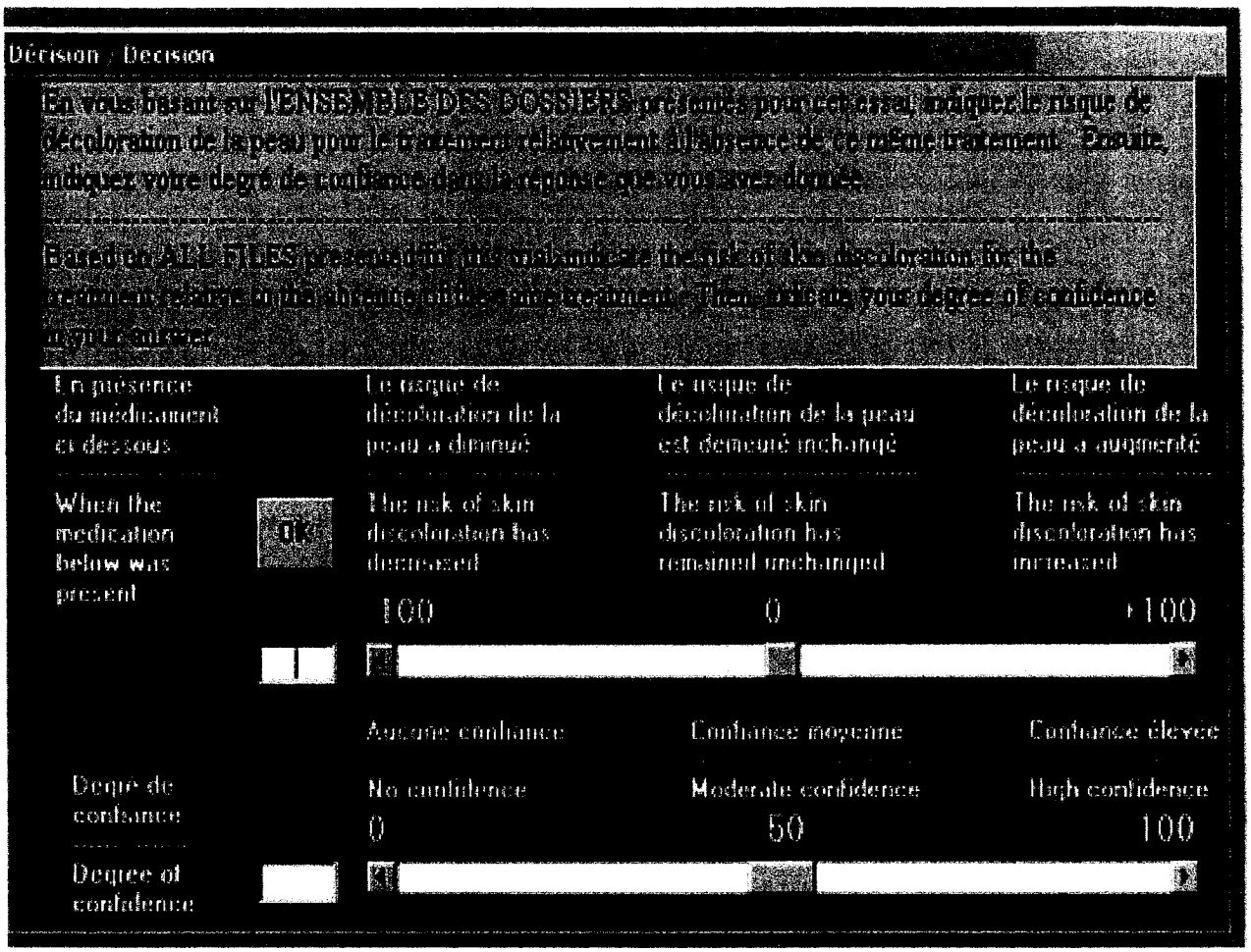

Figure 5: Contingency Judgment and Confidence Rating Screen

After the practice conditions were completed, the same procedure was carried out for each of 9 experimental sets of 40 trials. The values of ΔP for the nine sets were arranged to form a 3 X 3 design with ΔP's of .1, .2, or .3 repeated three times each according to three power patterns, low ascending, high fixed and descending, as detailed in Table 3. The order of presentation of the experimental conditions was randomized for each participant.

Table 3: Design and Theoretical Predictions for Experiment 1

a	*b*	*c*	*d*	*Total*	*P(e|i)*	*P(e|~i)*	*ΔP*	p_i	*Rel.*[1]
Practice conditions									
20	10	0	10	40	0.67	0.00	0.67	0.67	1.00
0	10	20	10	40	0.00	0.67	-0.67	1.00	0.33
10	10	10	10	40	0.50	0.50	0.00	0.00	0.50
Low ascending power									
4	16	2	18	40	0.20	0.10	0.10	0.11	0.90
6	14	2	18	40	0.30	0.10	0.20	0.22	0.90
8	12	2	18	40	0.40	0.10	0.30	0.33	0.90
High fixed power									
20	0	18	2	40	1.00	0.90	0.10	1.00	0.10
20	0	16	4	40	1.00	0.80	0.20	1.00	0.20
20	0	14	6	40	1.00	0.70	0.30	1.00	0.30
Descending power									
20	0	18	2	40	1.00	0.90	0.10	1.00	0.10
16	4	12	8	40	0.80	0.60	0.20	0.50	0.40
6	14	0	20	40	0.30	0.00	0.30	0.30	1.00

[1]Reliability = $[(a+b)-(a+b)*P(e|\sim i)]/(a+b)$

For each experimental condition listed in Table 3, the same judgment and confidence screen was presented once the 40 patients' records had been shown. Participants were free to take as much time as needed to decide on their responses, but once they clicked to continue to the next condition, their responses could not be changed. After the participants were finished with this phase, there was a reminder that a new series of trials was about to begin and that they should disregard all previous trials when considering this next one.

Statistical analyses

An alpha level of .05 was selected as the criterion for tests of significance. Univariate *F* values are reported throughout. All within-subject effects were corroborated with multivariate tests. Post-hoc comparisons used Tukey's test with the appropriate *MSE* from the *ANOVA*. While the Low-Medium-High versus Low-High contingency factors do not necessarily represent true separate factors, the gross versus fine gradations in contingency formed the basis of the analytical hypotheses *a priori*. As such, a factorial versus one-way *ANOVA* was employed.

Results

The causality judgments are depicted in Figure 6 as a function of actual contingency. Expected values based on the Power PC theory are presented for comparison.

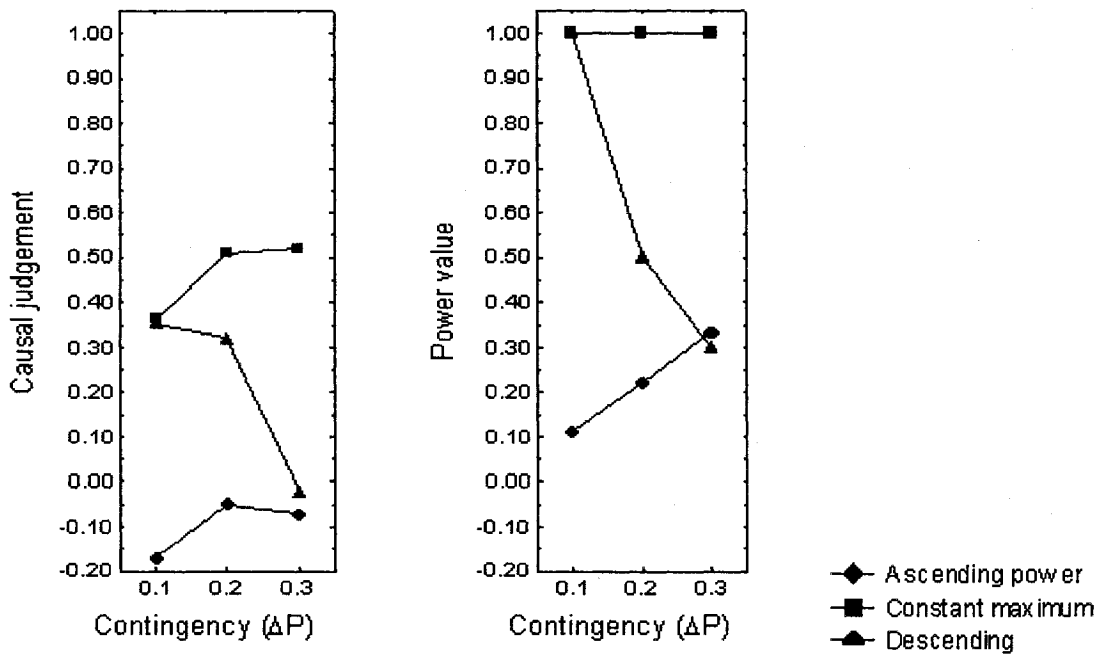

Figure 6: Obtained Causal Judgments (left) Versus Predicted Power PC Values (right)

The first impression gained from the figure is that the overall pattern of obtained judgments was remarkably similar to the pattern of predictions based on the power PC theory. The slopes for each of the three power-related experimental conditions are ascending, constant or descending as expected. Yet, at the same time the obtained judgments appear to fall considerably lower than the predicted values on the same judgment scale. The interpretation of this observation will be continued below.

The judgments were analysed using a repeated-measures analysis of variance with a 3 (contingency) x 3 (power) factorial design. This analysis revealed a significant main effect of power, $F(2, 46) = 21.32$, $MSE = 2672.56$ and an interaction of power by contingency, $F(4, 92) = 4.96$, $p < .05$, due to the differing slopes of power conditions. The lack of a main effect of contingency indicates that the three chosen levels of ΔP were not easily discriminated by the participants. This could be due to the fact that all three chosen ΔP values were relatively small (.1, .2, .3); yet other experiments (Shanks, 1985a; Valée-Tourangeau et al, 1997, 1998; Lober & Shanks, 2000) have demonstrated participants' ability to discriminate changes in ΔP more subtle than .1. Another possibility is that the lack of contingency discrimination in the present experiment may be due to an averaging effect brought about by the within subject design. Each ΔP level is tested at a relatively low, medium and high level of power. The mean power values, p_i, for $\Delta P = .1, .2, .3$ are .70, .57, and .54 respectively across the experimental conditions. If judgments are somehow affected by both contingency and power related variables, then the mixture may have been very hard to discriminate.

It should be noted that while the overall pattern of the contingency estimates has the shape predicted by the Power PC theory, all the values are much lower than expected. Of the four conditions that were predicted to reach maximal power (1.0), only two are slightly above .5 on the contingency judgment scale. The judgments were not statistically compared with predicted values because there may be scaling differences between the subjective and objective measurements. However, at first glance it appears that participants' estimates never approach what would be considered a 'maximal' value.

For the confidence data, the overall pattern is shown in Figure 7. The repeated-measures ANOVA for confidence, using the same 3 (contingency) x 3 (power) design, revealed only a main effect of power, F (2, 46) = 7.00, confirming that participants did discriminate among some aspects of the experimental conditions after all. Post-hoc testing located the significant difference between the constant and ascending power conditions. Within the design of this first experiment, it seems that confidence is not directly related to ΔP.

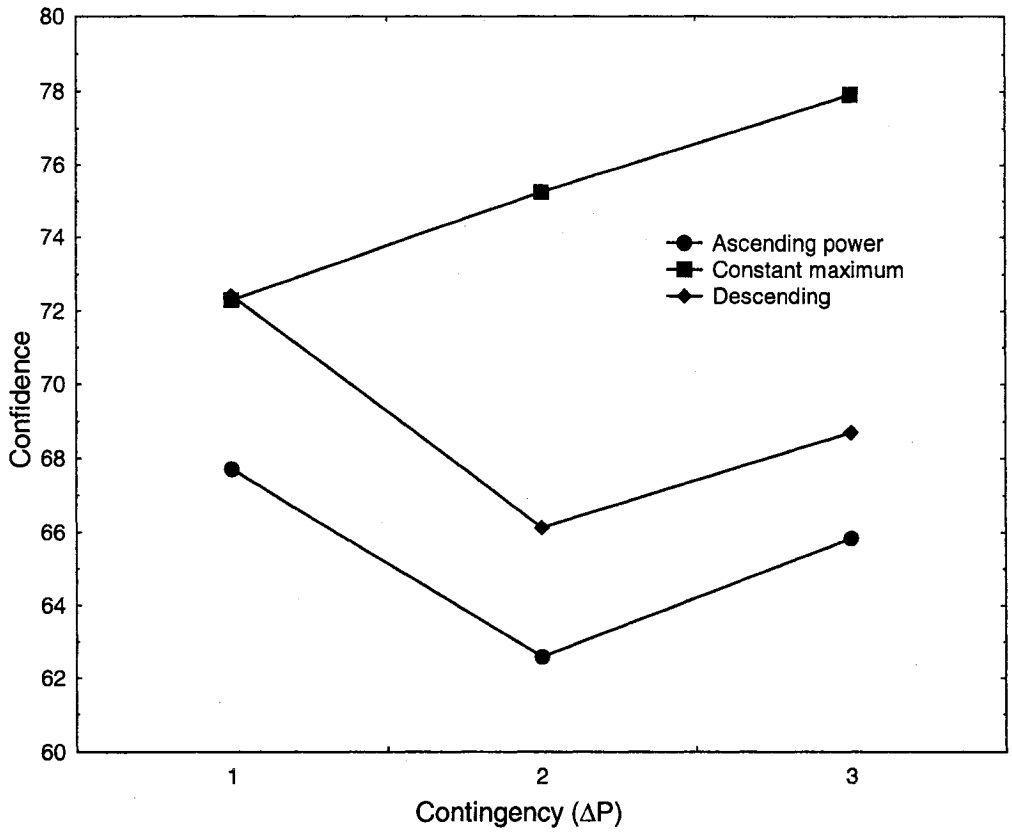

Figure 7: Confidence Ratings for Experiment 1

Although there is a degree of relation between power and confidence, the link is not completely systematic. There may be a yet undiscovered underlying variable. Whatever it is, it cannot be reliability as this variable was at a constant high in the low ascending power condition, where confidence was at its lowest. This suggests that further research is warranted in examining the relationship between confidence, reliability and contingency judgment. One possible alternative interpretation of this issue has been put forward by Clément, Mercier and Pastò (2002), who consider the role that complexity or contrast (distribution of *a-b-c-d* cells) plays in confidence. They argue that experimental conditions which have low complexity and high contrast, will allow for the research participants to be more confident in their judgments. Based on the data available from the

present experiment, the four conditions where power is maximal also happen to have one empty cell (low complexity / high contrast) and they all receive the highest confidence scores. It will be worth noting whether this pattern holds true within the parameters of the next experiments.

While the participants' contingency judgments in this first experiment seem to fall short of true 'maximal' values as would be predicted by the Power PC theory, the evidence at hand remains inconclusive. It may be possible that the highest values reported by these participants indeed represent their subjective maximal causal strength. On the other hand, if it can be demonstrated that causal strength can indeed be higher for other conditions of $p = 1.0$, we would need to revisit whether or not the highest values here actually represent maximal judgments, and the Power PC theory would fall short as a complete explanation of human contingency judgment.

Experiment 2

Experiment 1 showed that while the *pattern* of the Power PC theory's predictions held true, the obtained results did not seem to approach true *maximal* judgments. This second experiment was designed to test whether participants' judgments could indeed approach maximal values in at least some of the conditions in which the Power PC theory predicts that they will. The conditions chosen for this experiment allow for variation of ΔP while causal power maintained a value of 1.0. The Power PC theory predicts that results in all conditions will be maximal. However, based on Experiment 1, we know that judgments are sometimes low even when power is maximal. Since all power conditions in Experiment 1 were crossed with low contingencies, thus potentially producing some

sort of averaging effect between the conflicting information of power and contingency, Experiment 2 studied this by systematically crossing power with low or high contingencies.

Method

Participants

Twenty-four volunteer undergraduate students (12 Male, 12 Female) from the University of Ottawa took part in the experiment. 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. 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 icon measured approximately 9mm high and 5mm wide, while the face icon had a diameter of approximately 12mm. 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. Each of the four possible combinations was represented in Figures 1 through 4 as in Experiment 1. The treatment

(pill) and outcome (skin condition) were presented simultaneously, for the duration of 1 second, with an inter-stimulus interval of 1 second. The two pill colours – blue and red – were counterbalanced according to odd and even numbered experimental conditions. The colours were only used to increase the generality of the results but were not included in the analysis as there was no theoretical hypothesis attached to them.

Procedure

Participants were seated in front of the computer and oriented to the experiment with written instructions which appeared on the computer's screen. At the same time, icons representing each of the possible conditions (presence/absence of medication, and presence/absence of skin condition) were shown to familiarize the participants with the stimuli. These initial instructions were the same as in Experiment 1.

The three practice conditions followed, consisting of 40 medical files each. As listed in the top portion of Table 3, the first practice had a ΔP of .67, the second -.67 and the third was 0. The order of the practice conditions was fixed and chosen to emphasize contrast in covariation and in power (.67, 1, and 0) simultaneously. Each condition included the presentation of a set of 40 fictitious medical files, each of which constituted a trial. Each set of 40 trials was made up of two blocks of 20 trials keeping covariation and other variables constant. The order of presentation of the trials was randomized within each block of each condition. At the end of each condition, participants were asked to judge the relative efficacy of the treatment on a scale from -100 to 100 and rate their confidence in the judgment they just made on a scale from 0 to 100 as in Experiment 1. Participants were not given feedback regarding the accuracy of their judgments at any point in order to genuinely capture their subjective impression rather

than shape their judgement. Nevertheless, the normative contingencies used in the practice conditions were chosen to expose each participant to a strong positive, a strong negative, and a null contingency before proceeding to the experimental trials. 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.

The experimental conditions are presented in Table 4 below. Note that the order of presentation of both conditions and trials was randomized for each participant.

Table 4: Experiment 2 Design and Theoretical Predictions

a	***b***	***c***	***d***	***P(e|i)***	***P(e|~i)***	ΔP	**Reliability**[1]	p_i
20	0	18	2	1.00	0.90	0.10	0.10	**1.00**
20	0	16	4	1.00	0.80	0.20	0.20	**1.00**
20	0	14	6	1.00	0.70	0.30	0.30	**1.00**
20	0	4	16	1.00	0.20	0.80	0.80	**1.00**
20	0	2	18	1.00	0.10	0.90	0.90	**1.00**
20	0	0	20	1.00	0.00	1.00	1.00	**1.00**

[1] Reliability: $[(a+b)-(a+b)*P(e|\sim i)]/(a+b)$

In order to test our hypothesis, three low-contingency conditions were used (ΔP = .1, .2, .3 with $P(e|\sim i)$ = .9, .8, .7 respectively), along with three high-contingency conditions (ΔP = .8, .9, 1.0 with $P(e|\sim i)$ = .2, .1, .0 respectively) which all generated the same prediction of p_i = 1 from the Power PC theory. Thus two global levels of Low vs. High ΔP were crossed with three specific values of smallest, intermediate and highest strength within each of those conditions.

Results

From the contingency judgment results which are depicted in Figure 8, it appears that participants' judgments were more correlated with contingency than power. For all six experimental conditions, the Power PC theory predicts a constant maximal judgment, yet there is a distinct gap between the low and high ΔP conditions. However, within each of the ΔP conditions, participants seem to not be particularly sensitive to the increasing pattern of contingency.

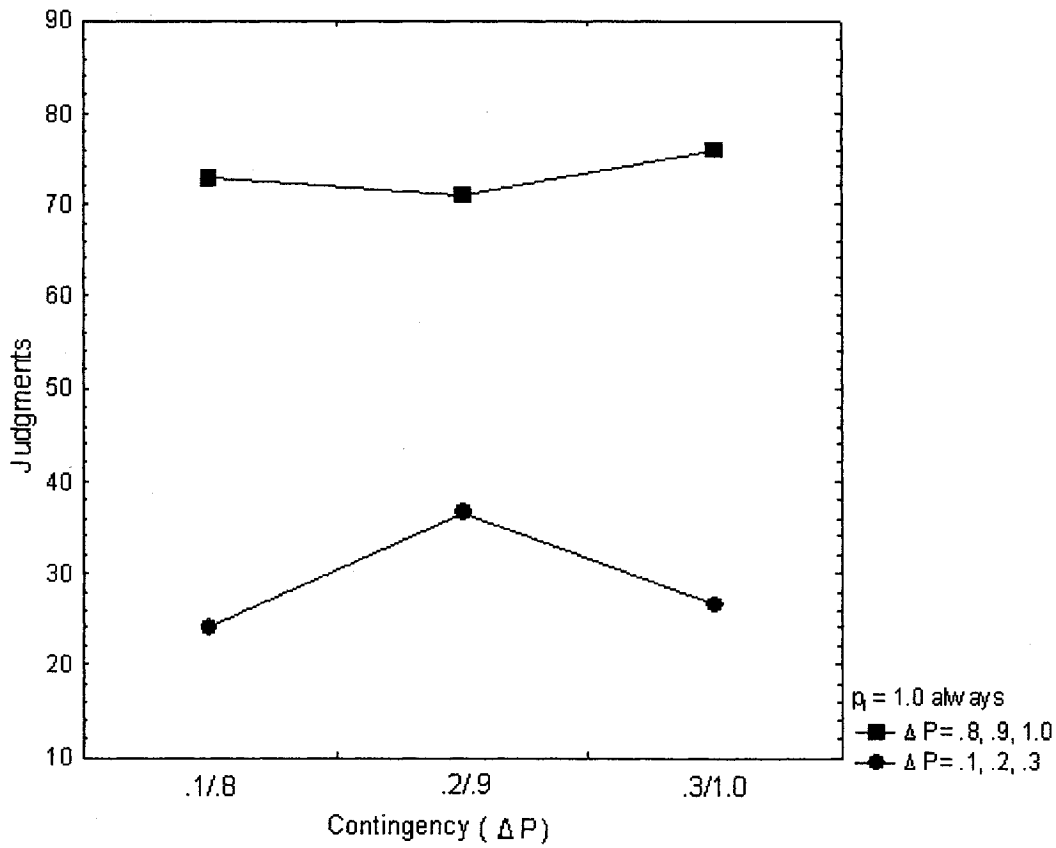

Figure 8: Obtained Causal Judgments for Experiment 2

The data were analyzed according to a within-subject factorial analysis of variance with two main levels of contingency (high, low) crossed with three sub-levels (high, medium, low). A main effect was obtained for the Low-High ΔP contrast, F (1, 23) = 54.4, MSE = 1605, with no significant effect of the individual ΔP values within those categories. This demonstrates that the participants were able to discriminate between the low and high ΔP conditions for which Power PC predicted equal levels of causal power. This aspect of the judgments is therefore inconsistent with the Power PC theory. This reinforces the conclusion from the first experiment that while the *pattern* expected by the Power PC theory was obtained, the level of the judgments cannot be dismissed as merely a scaling issue between participants' subjective judgments and the normative predictions of the Power PC. At the same time, there is only weak support for the basic PCM. The ability of participants to discern between the low and high ΔP conditions is consistent with the PCM, but there was no observed increase in judgments between the individual .1, .2, .3 or .7, .8, .9 ΔP conditions.

As shown in Figure 9, confidence ratings were significantly lower for conditions where covariation and power predictions were the most contradictory (ΔP / power: .1 / 1, .2 / 1 and .3 / 1), F (1, 23) = 21.82, MSE = 262.93, no other significant effect in the factorial ANOVA). Perhaps it is this variable rather than reliability that affects confidence the most.

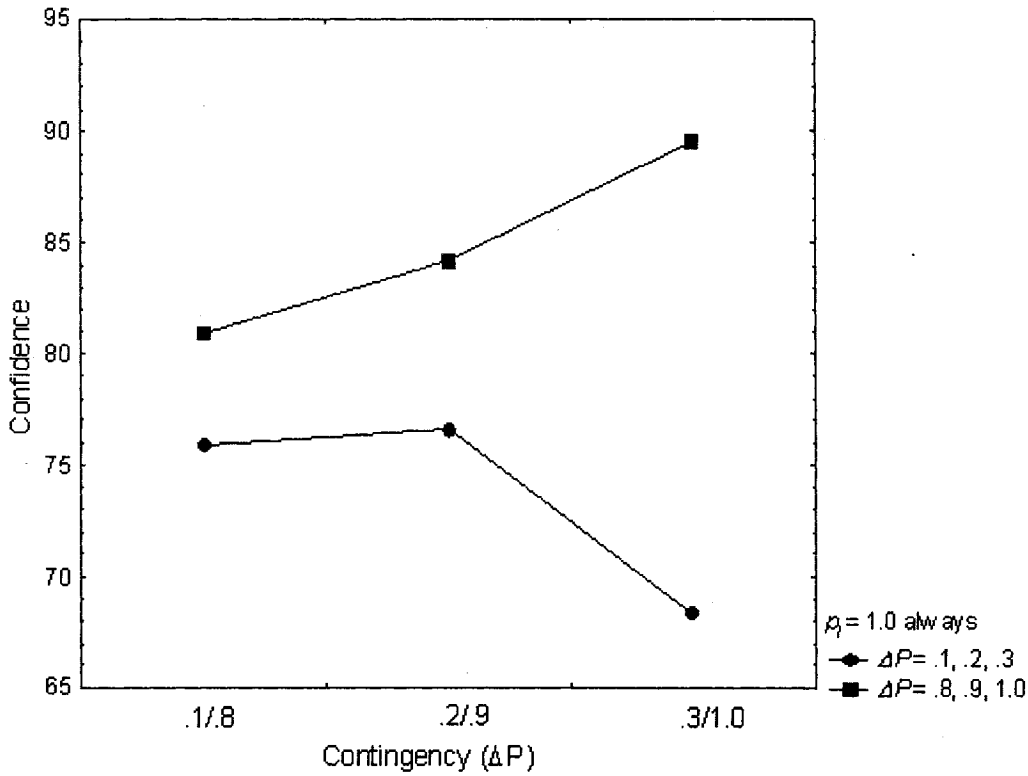

Figure 9: Confidence Ratings for Experiment 2

While confidence is not a simple function of reliability, it must be noted that in all the conditions of Experiments 1 and 2 where power should have been judged maximal (p_i = 1.0), reliability was somewhat low ($.1 \leq R \leq .51$). Thus, although reliability was not the main source of influence on confidence when this variable was measured directly, it may have affected judgment in some other way that remains to be determined.

In order to explore the influence of reliability on judgments, a third experiment was carried out.

Experiment 3

This experiment was designed to investigate the roles of reliability, power and ΔP in contingency judgment, following the potentially confounded results of the second

experiment. If participants' judgments were found to vary systematically with reliability, the results of our first two experiments would need to be revisited.

To summarize the nature of the potential confound, reliability can be seen as an attenuating factor which would reduce the potential for participants' causality judgments to reach their peak. Buehner and Cheng (1997) and Buehner, Cheng and Clifford (2003) explain reliability as the number of opportunities the candidate cause (i) has to demonstrate its effect (e), minus the number of these trials on which the participant might expect the effect to happen simply by chance considering the base rate of the effect ($P(e|\sim i)$). That is to say that when the base rate of the effect is high, participants need to witness many trials on which the effect is demonstrated in order to assess the causal power of the candidate cause. This would be a condition of low reliability. When the base rate of the effect is low or zero, almost all or all of the occurrences of the effect can be taken into consideration as evidence of the causal power of the candidate cause. This would be a condition of high reliability. From the Power PC standpoint alone, all six experimental conditions listed in Experiment 2 ought to have produced similar, maximal causal judgments. The obtained result, which demonstrated a significant effect of ΔP (and hence, the Power PC prediction of equal judgments fell flat), could also have been argued to be due to the lower reliability for those conditions in which participants' judgments did not appear to reach maximal levels. Addressing this possibility will be the fundamental purpose of this third experiment.

Method

Participants

Twenty-four volunteer undergraduate students (12 Male, 12 Female) from the University of Ottawa took part in the experiment. 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. The experimental task was programmed with Microsoft Visual Basic Professional 6.0.

Stimuli

The stimuli were as in Experiment 1.

Procedure

Participants were seated in front of the computer and oriented to the experiment as in the previous two experiments. The three practice conditions with ΔP of .67, -.67 and 0 then followed. Instructions were as before.

The experimental conditions are presented in Table 5. Once again the order of presentation of the conditions and trials was randomized for each participant.

Table 5: Design of Experiment 3

a	*b*	*a*+*b*	*c*	*d*	*c*+*d*	Total	**ProbC**	$P(e\|i)$	$P(e\|\sim i)$	ΔP	***pi***	**Reliability**
9	1	10	27	3	30	40	**0.25**	0.90	0.90	0.00	**0.00**	**0.10**
0	10	10	0	30	30	40	**0.25**	0.00	0.00	0.00	**0.00**	**1.00**
10	0	10	27	3	30	40	**0.25**	1.00	0.90	0.10	**1.00**	**0.10**
10	0	10	0	30	30	40	**0.25**	1.00	0.00	1.00	**1.00**	**1.00**
18	2	20	18	2	20	40	**0.50**	0.90	0.90	0.00	**0.00**	**0.10**
0	20	20	0	20	20	40	**0.50**	0.00	0.00	0.00	**0.00**	**1.00**
20	0	20	18	2	20	40	**0.50**	1.00	0.90	0.10	**1.00**	**0.10**
20	0	20	0	20	20	40	**0.50**	1.00	0.00	1.00	**1.00**	**1.00**
27	3	30	9	1	10	40	**0.75**	0.90	0.90	0.00	**0.00**	**0.10**
0	30	30	0	10	10	40	**0.75**	0.00	0.00	0.00	**0.00**	**1.00**
30	0	30	9	1	10	40	**0.75**	1.00	0.90	0.10	**1.00**	**0.10**
30	0	30	0	10	10	40	**0.75**	1.00	0.00	1.00	**1.00**	**1.00**

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+Reliability model, 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 maximize the generalizability of any results from the two previous experiments where this variable was set to 1.0 whenever power was 1.0. A 3 x 2 x 2 (Low-Med-High probability of the cause, by Low-High power, and Low-High reliability within each of those conditions) factorial design was used.

Results

Figure 10 shows 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. Although power was sometimes judged maximal when its predicted value was 1.0, it was also judged consistently lower when ΔP and reliability were low. That fact alone precludes a firm conclusion on whether the dampened causal perception resulted from low reliability or from the conflictual information between power and contingency (ΔP). However, the accompanying effect of reliability in further reducing judgements even below 0 while power and contingency both remained high suggests a possible key role played by this variable.

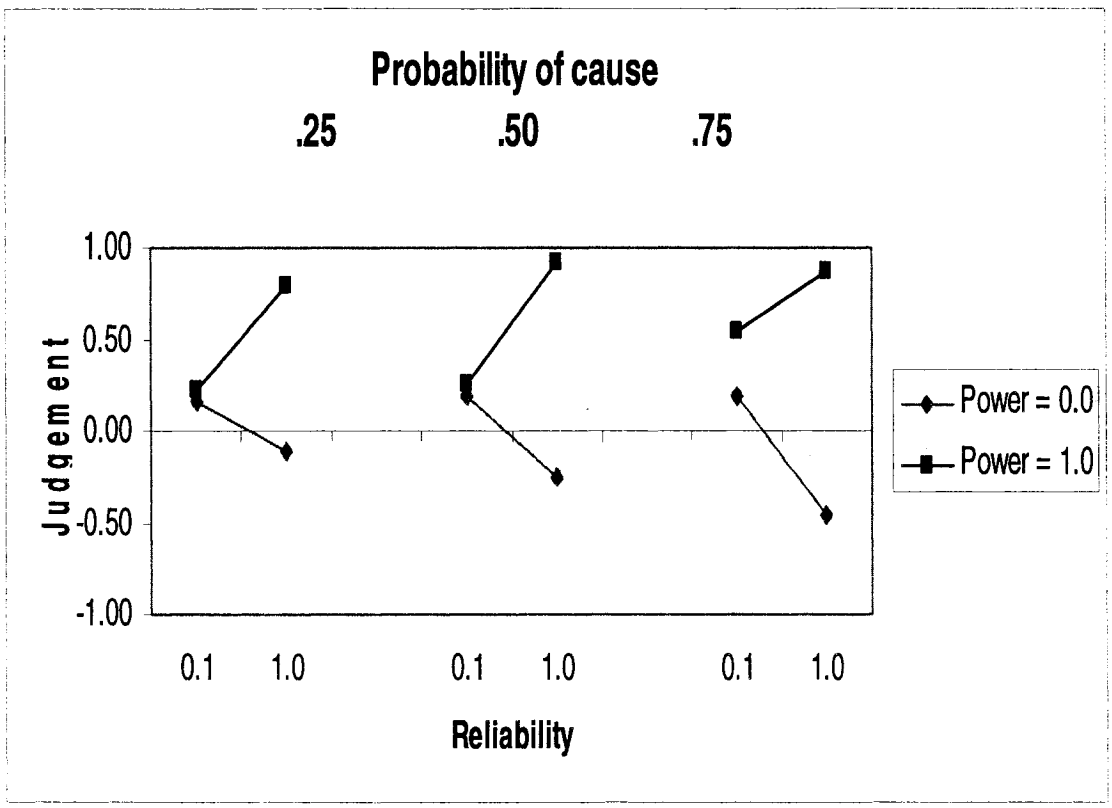

Figure 10: Judgments for Experiment 3

These observations are corroborated by the statistical analyses that found a significant main effect of power ($F_{1,23} = 111$) and a significant interaction of power and probability of cause ($F_{2,22} = 5.44$). Power also interacts significantly with reliability ($F_{1,23} = 105$). Finally, there was also a significant interaction between probability of cause and reliability ($F_{2,22} = 4.42, p = .05$).

Confidence results, depicted in Figure 11, showed that confidence is influenced by power ($F_{1,23} = 15.2$), reliability ($F_{1,23} = 43.6$) and the probability of cause ($F_{3,46} = 3.69$), although the latter two interact ($F_{2,46} = 7.20$) such that confidence is less reduced by low reliability at higher levels of the probability of cause. In this manner, participant confidence is bolstered by higher overall probability of the cause, coupled with high reliability.

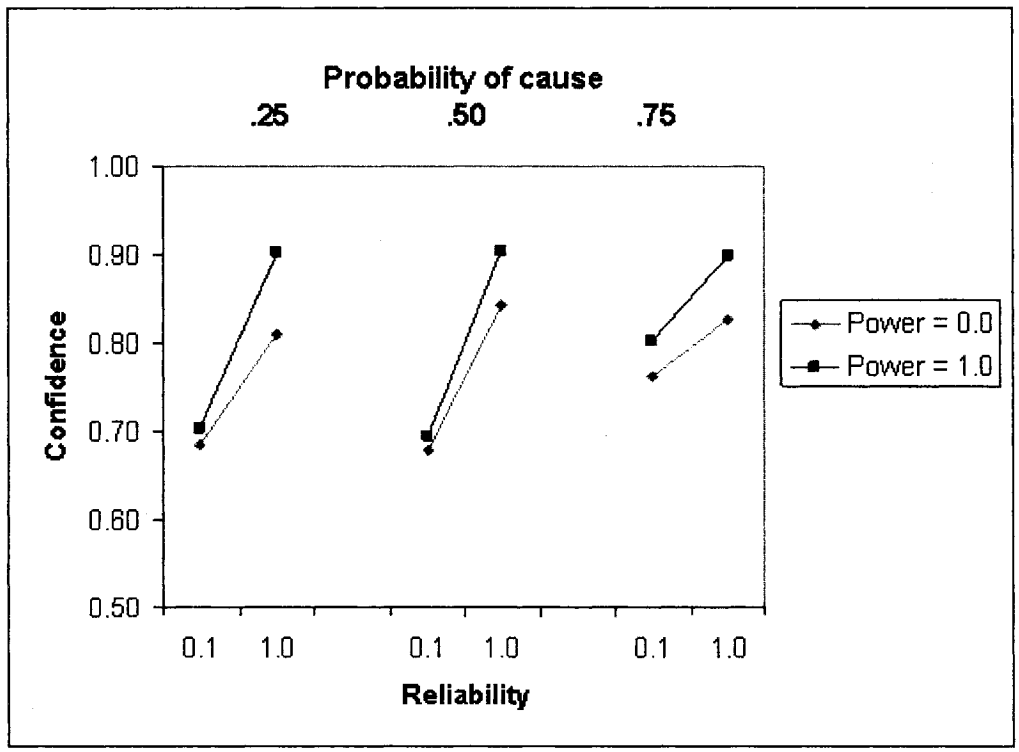

Figure 11: Confidence Results for Experiment 3

Although the significant interaction between reliability and power appears to support Buehner's position that judgments will always be lower when reliability is low, this falls short of the entire story. Recall that judgments were less than maximal in spite of combined high power and high reliability in Experiment 1. Secondly, reliability does not appear to work solely through confidence as hypothesized. Third, a combined power-reliability concept does not capture effects such as the probability of cause in Experiment 3. These issues will be explored further in the general discussion.

Experiment 4

With the results of the previous three experiments demonstrating the insufficiency of the Power PC theory in terms of generative causal power (positive contingencies), a proposed fourth experiment was designed within the realm of preventive causal power (negative contingencies). This goal of this experiment was to determine whether the findings of the first three experiments with regards to the Power PC theory, reliability, and confidence, will hold true in conditions of negative contingency. If these findings are indeed replicated, Power PC and any subsequent interaction of Power PC and reliability will both have been shown to be insufficient models of contingency judgment. At the same time, the lessons learned about the respective roles of confidence and reliability in human contingency judgment will be further validated.

Method

Participants

Twenty-four volunteer undergraduate students (8 Male, 16 Female) from the University of Ottawa took part in the experiment. 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. The experimental task was programmed with Microsoft Visual Basic Professional 6.0.

Stimuli

The stimuli were as in the previous three experiments.

Procedure

Participants were seated in front of the computer and oriented to the experiment as before. The same three practice conditions were followed by the experimental conditions detailed in Table 7.

Table 7: Design of Experiment 4

a	*b*	*c*	*d*	*P(e|i)*	*P(e|~i)*	*ΔP*	*pi*	**Reliability**
20	0	18	2	1.00	0.90	.1	1.00	0.10
16	4	12	8	.80	0.60	.2	.50	0.40
6	14	0	20	.30	0.00	.3	.30	1.00
20	0	6	14	1.0	.3	.7	1.0	0.70
20	0	4	16	1.0	.2	.8	1.0	0.80
20	0	2	18	1.0	.1	.9	1.0	0.90
0	20	2	18	.00	.10	-.10	1.0	0.10
4	16	8	12	.20	.40	-.20	.50	0.40
14	6	20	0	.70	1.0	-.30	.30	1.00
0	20	14	6	.00	.70	-.70	1.0	0.70
0	20	16	4	.00	.80	-.80	1.0	0.80
0	20	18	2	.00	.90	-.90	1.0	0.90

Six conditions of generative causal power (positive contingencies) were mirrored with six conditions of preventive power (negative contingencies). The specific cell values for each given trial were chosen to best replicate the counter-intuitive predictions of the Power PC theory in the negative contingency matrix. That is, there are both positive and negative contingency conditions in which Power increases as ΔP weakens, as well as conditions in which power remains maximal as ΔP changes.

Since the first three experiments did not include any negative ΔP values, the original formulation of reliability (Equations 7 and 8) was used. However, a modification to this formulation is required when examining preventive causal power. While it has not been explicitly specified in the literature, Buehner (personal communication, October 5, 2005) did confirm how reliability is to be calculated for negative contingencies. It follows logically that the reliability of a preventive cause will be based on the formula given in Equations 10 and 11 rather than the original formula. Essentially the change from $P(e|\sim i)$ to $1\text{-}P(e|\sim i)$ indicates that the inverse of the base rate of the effect should be used. This is because with a preventive causal power, the base rate we are interested in is how often the effect *fails to occur* in the absence of the presumed preventive cause, which is the inverse of the likelihood of it occurring. With this formulation of reliability in negative contingency space, the fourth experiment had a secondary aim to determine the relationship between reliability and causal judgments in negative contingencies.

$$R = \left(\frac{(a+b)-(a+b)(1-p(e|\sim i))}{(a+b)} \right) \tag{10}$$

$$R = \left(\frac{(a+b) - (a+b)\left(1 - \frac{c}{c+d}\right)}{(a+b)} \right) \tag{11}$$

Results

Before analyzing the data, judgements obtained under preventive conditions only were recoded to be consistent with the scale used for the generative contingencies by inverting their sign. Thus a reported preventive contingency of -50 would have been recoded as +50, and a value of +5 would have been re-coded as -5. This was done to avoid spurious statistical interactions solely due to sign. We are more interested in whether judging a negative contingency produces any different pattern than judging a positive contingency, than in whether participants can simply perceive negative contingencies.

Analysis of the data from Experiment 4 was carried out using a 3 (Low-Med-High ΔP) x 2 (Low-High ΔP) x 2 (Generative-Preventive Cause) ANOVA. Figure 12 below presents the results graphically, for the respective conditions of generative contingency and preventive contingency. There was no significant difference between generative and recoded preventive conditions. Thus there is no reason to believe the conclusions drawn from the first three experiments do not logically apply equally well to negative contingency space. Main effects of Low-High ΔP ($F_{1,23} = 209$) and Low-Med-High ΔP ($F_{2,46} = 14.4$) were found, as well as a significant interaction of the two ($F_{2,46} = 18.4$). This reveals that participants were capable of discriminating between overall high and low ΔP values, as well as between the more subtle levels within both the high and low ΔP conditions. The interaction of the two seems to be caused by dramatically low causal

judgments for ΔP = +/- .30. Both of these conditions have low ΔP and p_i values, and high reliability. This seems to be inconsistent with both the ΔP and *PowerPC* patterns of contingency, particularly if p_i values are expected to be moderated by reliability as the value of reliability for these conditions is maximal and thus would be expected to raise the participants' judgments. It is noteworthy that these conditions also involved a base rate of zero (i.e., the effect never occurred in the absence of the candidate cause). However there appears to be no theoretical model which would account for such results. At this point, they remain anomalous.

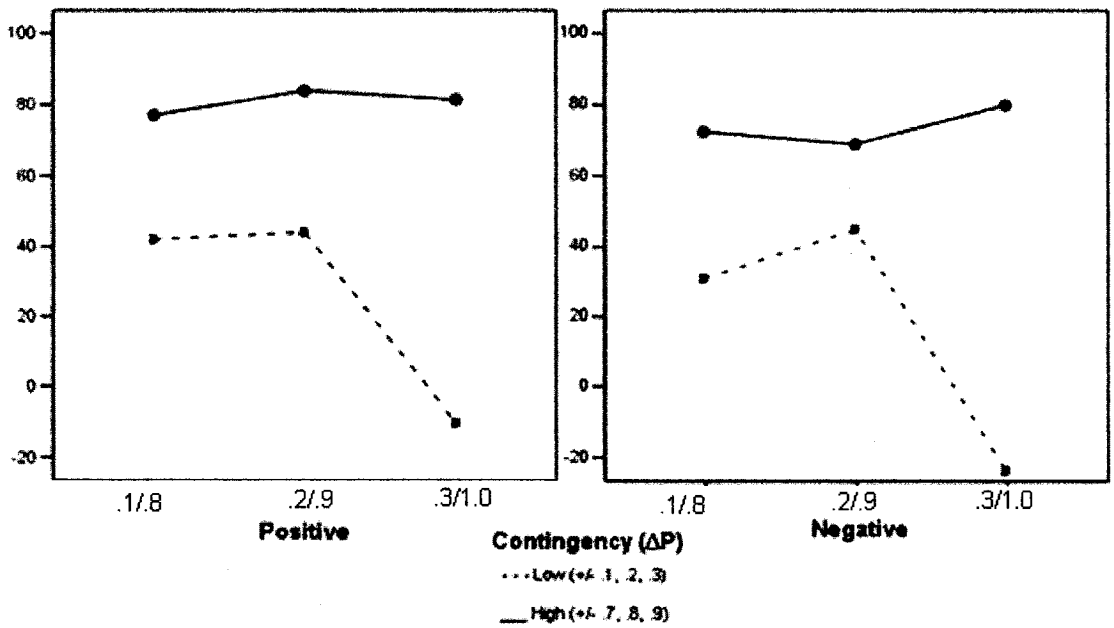

Figure 12: Judgments for Positive and Negative Contingencies

Confidence ratings provided by the participants on the same conditions, as presented in Figure 13, yielded a main effect of Low-High ΔP ($F_{1,23}$ = 4.05), but no main effect for Low-Med-High ΔP. A significant interaction of the two ΔP factors also

remained ($F_{2,46}$ = 10.0). The role that reliability plays in confidence is not immediately clear from these results, particularly as characterized by the descending confidence line for ΔP = -.1, -.2, -.3 while reliability is increasing to from .1 to .4 to 1.0. When the confidence results are directly paired to the respective judgments, it appears that the problematic ΔP = +/- .3 condition also generated lower confidence that most other conditions. But what constitutes a 'low' confidence score? With the results from this experiment, and given the reliability effect found in Experiment 3, the reliability-confidence relation warrants further attention. The nature of the interaction between reliability and Power PC, and further potential theoretical implications of reliability are discussed in the following section.

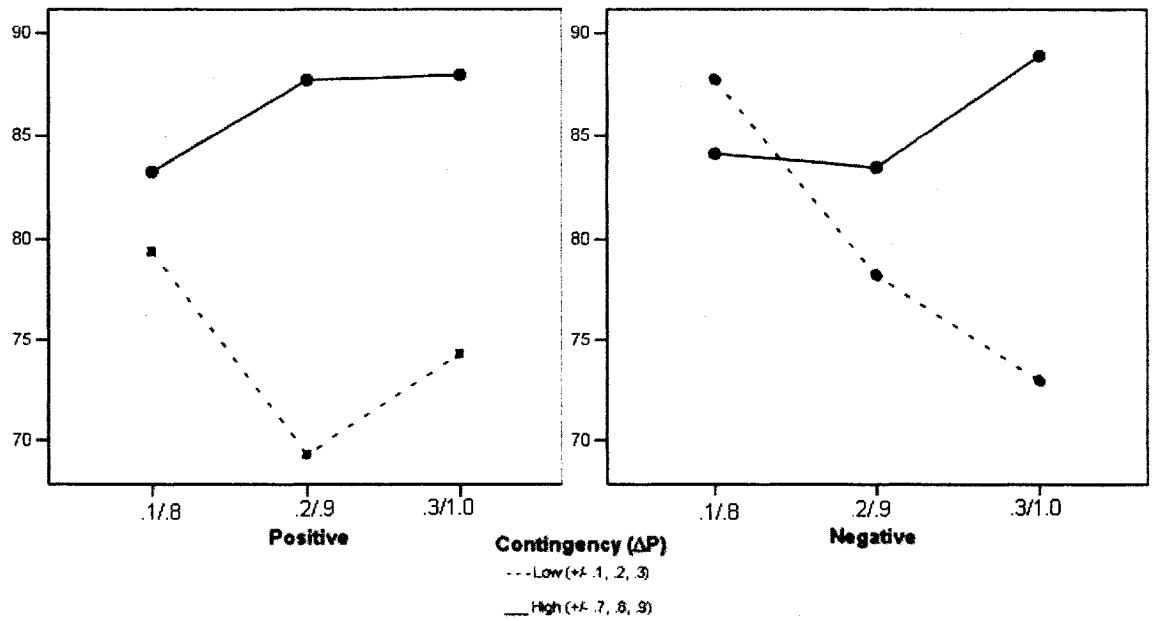

Figure 13: Confidence Ratings for Positive and Negative Contingencies

Chapter 3: Discussion

Summary of the experimental results

As a test of the validity of the Power PC theory, Experiment 1 returned results which were consistent with the model in pattern, but not in terms of absolute levels. The similarity in pattern between the predicted Power PC results and the actual results was striking and even unexpected, given how the conditions chosen for Experiment 1 were specifically those for which the Power PC theory produced the most counter-intuitive predictions (i.e., predictions of maximal and constant causal power in light of minimal or varying levels of underlying ΔP). The fact that the absolute judgment levels did not reach values of 100, as would be expected from the normative p values of 1.0, was still consistent with the Power PC theory as long as participants' judgments were at their own subjective maximum. To reiterate an earlier explanation of this reasoning, we cannot assume that a participant response of 87 or 92 is categorically different from the normative, predicted value of 100. The normative and obtained judgments (objective and subjective, respectively) may be operating on different scales of measurement, and it is not the 'accuracy' of the judgments which is being evaluated, but whether the pattern of judgments is consistent with one theory's predictions or another. But this remained an unverified question – did the judgment levels obtained in experiment 1 for conditions with p values of 1.0 represent maximal judgments?

This question was examined in Experiment 2, and an affirmative response would provide yet further support for the Power PC theory. However, Experiment 2 provided clear evidence that for conditions of equal predicted causal power ($p_i = 1.0$ for all conditions), there were significant variations, more closely tied to the pattern of ΔP

values. This would normally be sufficient to dismiss the Power PC argument of a ceiling effect accounting for the lower judgment levels in Experiment 1; however, proponents of the Power PC theory had argued that participants' judgments may be tempered by reliability (Buehner & Cheng, 1997; Buehner, Cheng & Clifford, 2003). Indeed, from the design of Experiment 2 (see Table 4), any obtained result consistent with the normative ΔP values would also be consistent with reliability. Thus, the results may not actually reflect true differential causal perceptions but, rather, the establishment of a reliability effect.

In order to discern between these two alternative interpretations, Experiment 3 was designed to simultaneously control for all three of ΔP, p_i, and R. Low and high values of each were systematically crossed with low and high values of the other two parameters. The overall probability of the cause was also controlled, in order to rule it out as a confound with p_i, as it had always been equal to 1 whenever $p_i = 1$ in the earlier experiments. The results from this experiment revealed an interaction of power and reliability, as had been argued by Buehner and Cheng (1997) and Buehner, Cheng, and Clifford (2003). There was no main effect of probability of the cause, ruling out another potential confound from the earlier experiments. Additionally, reliability had a significant effect on confidence, which is also consistent with the argument that low reliability could have been an explanation of low judgments in the first two experiments. While these effects and interactions remain, it will be shown in the next section the conjunction of Power PC with reliability ultimately reduces to the original ΔP formulation. This is not meant to discredit the influence of reliability on confidence or judgments, but merely that reliability cannot be seen as a feasible adjunct to the original p_i formula.

Experiment 4 was carried out to determine if the effects noted in the earlier experiments could be said to apply to negative contingency space. In this regard, there was no main effect of positive versus negative ΔP, indicating that the factors influencing preventive causal induction are the same as those influencing generative causal induction. Participants were able to discriminate between both gross and fine gradations in ΔP, but two outlier conditions produced results inconsistent with either p_i, ΔP, or reliability. These conditions had zero trials where the effect was present in the absence of the cause, which may have had a lowering effect on participants' judgments, but is currently not understood from any particular theoretical standpoint. Confidence scores for these conditions were also among the lowest for the entire experiment.

Reliability as an Adjunct to Power PC?

The first two of the experiments presented here were designed without including the concept of reliability. By revisiting the results of those experiments, it was clear that reliability was a plausible concept, and not including it was a potential confounding factor. As such, experiment three was designed to take reliability into account. Before discussing the results and implications of all four experiments, it is important to fully discuss the issue of a potential confounding factor – something which could dramatically alter the interpretation of the experimental results. Once this factor has been addressed, the following section returns to the consideration of the results.

Buehner et al. (2003) presented data from a series of four experiments investigating the roles of $P(e|{\sim}i)$ and ΔP on perceived causal power. In the first experiment, results from non-contingent conditions (i.e., $\Delta P=0$) deviated significantly from predicted causal power values of zero, as well as conditions where ΔP varied but

constant power was predicted. These variations from the Power PC theory were discussed as being possibly linked to ambiguity in the rating scale used, which led participants to "conflate reliability with causal strength" (2003, p. 1126). The ensuing experiments used a different cover story and procedure which "allowed for only one interpretation of the causal question" (2003, p. 1138), and the researchers report that the results from those experiments are consistent with Power PC.

Buehner and Cheng's (1997) and Buehner et al.'s (2003) reports unfortunately do not include a detailed formulation of what reliability is and how it can be computed. It seems clear that as ambiguity in the experimental task increases, the authors feel participants pay increased attention to reliability. From their discussion, reliability can be viewed conceptually 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 ($P(e|\sim i)$). In terms of alternate causes (a), reliability represents a correction of the number of (e) outcomes due to (i) based on the number of times one could expect (a) to have brought about the effect (e). This is the formulation that was represented in Equation 8 and 10 above for generative and preventive causes. These were confirmed by M. Buehner (personal communication, October 5, 2005) as his intended formulation.

While the combined results of experiments 1 and 2 from the current study appear to provide solid evidence that participants' judgments are not consistent with the Power PC theory, those experiments do not explore the concept of reliability. With reliability values uncontrolled in those two experiments, it became a potential confound which could explain the discrepancy between the obtained judgments and the Power PC

predictions. The third experiment then set out to introduce control over the reliability values for each condition, and answer the question of whether participants' judgments were deviating from p_i due to reliability. For this experiment, no factorial design was found in which reliability and power could be examined without losing experimental control over ΔP, so ΔP was not included in the ANOVA analysis. It proved difficult if not impossible to generate a complete factorial design controlling for each of these parameters (see Experiment 3 design). In the end, a main effect of reliability as well as an interaction of reliability and p_i were found. This appeared to support the argument that reliability had in fact been overlooked in the first two experiments, and presented a confounding factor. Indeed, the discovery of a statistical interaction between reliability and p_i seems to underline Buehner and Cheng's (1997) and Buehner's (2003) conception of how reliability and p_i work in conjunction (See equation 12 below). But why were the research participants sensitive to the interaction of these two factors? An answer to this question was not immediately clear, but a logical and empirical verification of their relation to each other was eventually discovered.

Lober and Shanks (2000) have briefly argued that the concept of reliability having an effect upon contingency judgments is invalid in certain conditions, by demonstrating that judgments in their experiments seem to have reached asymptote after approximately 20 trials. They found no evidence of an acquisition curve past this point and contend that participants' judgments in Buehner and Cheng (1997) had likely reached asymptote even after only 16 trials. Therefore, Lober and Shanks would argue that participants are not befuddled by any lack of reliability, as long as they are given sufficient trials on which to base their decisions. While this argument may be methodologically relevant and

theoretically sound, it stops short of investigating the concept completely. From the current research, it was initially suspected in trying to control for all three of ΔP, p_i, and R in the Experiment 3 design, that there seemed to be a systematic relation between all three constructs. After the experiment was run, the formulations of each of these parameters were examined and simplified to search for commonalities. The following equations demonstrate how a systematic relation exists not only conceptually, but is also computationally inevitable. If, as Buehner and Cheng (1997) and Buehner et al. (2003) hypothesize, research participants are indeed susceptible to conflating reliability with causal power, their judgment ratings would be expected to follow the predictions of a reliability-adjusted Power PC theory.

$$Power\ PC' = p_i * R \tag{12}$$

Re-expressing in terms of frequencies,

$$Power\ PC' = \left(\frac{\frac{a}{a+b} - \frac{c}{c+d}}{1 - \frac{c}{c+d}} \right) \left(\frac{(a+b) - (a+b)\left(\frac{c}{c+d}\right)}{(a+b)} \right) \tag{13}$$

Cancelling out the $(a+b)$ terms in numerator and denominator of the R portion of the preceding equation, it becomes:

$$Power\ PC' = \left(\frac{\frac{a}{a+b} - \frac{c}{c+d}}{1 - \frac{c}{c+d}} \right) \left(1 - \frac{c}{c+d} \right) \tag{14}$$

Further cancelling out the $1-c/(c+d)$ terms, the equation reduces to:

$$Power\ PC' = \left(\frac{a}{a+b} - \frac{c}{c+d} \right) \tag{15}$$

or

$$Power\ PC' = \Delta P \tag{16}$$

For preventive causes, a change in the formulation of reliability to incorporate 1-$P(e|{\sim}i)$ is necessary along with the appropriate expression of p_i. This was also confirmed with M. Buehner (Personal communication, October 5, 2005).

$$Power\ PC' = \left(\frac{-\Delta P}{p(e|{\sim}i)}\right)\left(\frac{(a+b)-(a+b)(1-p(e|{\sim}i))}{(a+b)}\right) \tag{17}$$

$$Power\ PC' = \left(\frac{-\left(\frac{a}{a+b}-\frac{c}{c+d}\right)}{\frac{c}{c+d}}\right)\left(\frac{(a+b)-(a+b)\left(1-\frac{c}{c+d}\right)}{(a+b)}\right) \tag{18}$$

$$Power\ PC' = \left(\frac{-\left(\frac{a}{a+b}-\frac{c}{c+d}\right)}{\frac{c}{c+d}}\right)\left(\frac{\left(\frac{c}{c+d}\right)}{1}\right) \tag{19}$$

$$Power\ PC' = -\left(\frac{a}{a+b}-\frac{c}{c+d}\right) \tag{20}$$

$$Power\ PC' = -\Delta P \tag{21}$$

Thus the demonstration of the relation between ΔP, p_i, and R applies to both generative and preventive formulations of Power PC. The computational proof of this relation reduces to a statement indicating that the conflation of causal power with reliability ultimately is equivalent to the original ΔP model of contingency. From that perspective, it is seen as an inadequate adjunct to the Power PC theory. Instead of acting

as an adjunct to the Power PC theory, the incorporation of reliability, as currently formulated, simply serves to return us to the most basic contingency model which has been proven inadequate in fully explaining human contingency judgment.

A question remains as to whether other formulations of reliability could capture the concept while avoiding circularity. For example, an alternate formulation of Power PC' could resemble Equation 22, where $f(R)$ represents some function applied to reliability. Such a function could possibly introduce a non-linear relation between p_i and participants judgments. Indeed, there are many situations in which such a formulation would no longer reduce directly to ΔP.

$$Power\ PC' = p_i * f(R) \tag{22}$$

Of course, the formulation in Equation 22 is highly speculative as the means by which reliability is said to modulate causal power is unclear in the absence of a more specific theoretical framework. It should also be noted that the function in Equation 22 is only applied to the R term and not the entire formulation. If it is required simply because of scaling needs for example then it might be required to apply to all the terms, including p_i. In that case, conflation of power with reliability would again reduce to a function of contingency as shown below.

$$Power\ PC' = f(pi * R) \tag{23}$$

$$f(\Delta P) = f(pi * R) \tag{24}$$

As such, any discussion of the nature of the interaction of reliability and p_i, could just as well begin with a discussion about ΔP. The experimental results which are inconsistent with the Power PC theory cannot be explained away by the incorporation of reliability. If that were the case, any conditions inconsistent with Power PC would necessarily be consistent with the PCM, and we have seen that that is not always the case.

The limitations of Power PC' considered here do not imply that reliability by itself has no role to play in causality judgement. The fact that this concept has been demonstrated to be fundamentally related to participants' judgments certainly brings up further questions. Do participants in other experiments also pay close attention to the proportion of opportunities the candidate cause (*i*) has to demonstrate its causal power over the effect (*e*), taking into account the proportion of those which could possibly be attributed to alternate cause (*a*)? If there is a specific linear or non-linear function by which reliability asserts its effect on contingency judgment, what is it? At this point it seems that reliability cannot be incorporated into the Power PC theory without the consequence of reducing Power PC' to ΔP, but the door has been opened for a new path of inquiry into the role of reliability, independent of the Power PC theory.

Alternative Approaches to Reliability and Contingency Judgment

This section will consider whether, on the basis of the four experiments presented here, there are any other viable roles for reliability to play. First, some of the potential candidates will be described, and then each will be evaluated quantitatively. Part of this process will involve considering whether reliability can form an adjunct to any of these alternatives. Utilizing a post-hoc approach to analysis, the questions answered will relate

to goodness-of-fit. As the experiments were not originally intended to incorporate these alternative approaches in controlled factorial design, the measure used here will be Pearson's *r* correlation rather than any form of ANOVA.

Griffiths and Tenenbaum (2005) employed a similar strategy using correlation measures to assess the post-hoc applicability of their causal support model to other researcher's data. Part of their argument is that other approaches to modeling contingency judgment ultimately only answer the question of causal strength, or "How strong is the causal relation between the candidate cause and effect?". Their causal support mechanism proposes to address the question of "Is there a causal relation between the candidate cause and effect?". Analogously, the difference between these questions is much like the difference between the statistical measures of effect size ("How strong is the relation?") and statistical significance ("Is there a significant relation?"). The model proposed by Griffiths and Tenenbaum (2005) is also intended to incorporate the notions of confidence and reliability in that it is sensitive to sample size. Both the PCM and Power PC approaches are fundamentally insensitive to sample size. Both are proportional measures, falling in a range of 0 and +/- 1. They are bound by the assumption that an adequate sample size has been reached for the participant to base their decision upon. Similar to the R-W approach, the causal support theory proposes that confidence and reliability will increase with a larger number of trials. In this respect, the causal support approach shares some characteristics with associative models.

Ultimately, the questions answered by Griffiths and Tenenbaum's (2005) causal support theory are not the same questions posed of the research participants in the four experiments presented here. Here, the participants were asked an 'effect size' type of

question: the degree to which the risk of skin discoloration has either increased or decreased. The experimental task was not one of determining whether a causal relation is present; one was assumed. For this reason, however promising the causal support approach may be, we cannot rationally form any conclusions about how well it addresses the data presented here. It will be important for further studies to modify the experimental task in order to evaluate the relative merit of the causal support theory.

Another approach to be considered here, but which has received relatively little experimental attention, is the Proportion of Confirming Instances (PCI) approach of White (1998, 2000, 2001, 2003, 2004). The PCI rule consists of an alternate-weighting of the PCM model. The logic behind the weighting system is that for generative causes, cells *a* and *d* represent 'confirmation' of the generative power of the cause, while cells *b* and *c* serve to disconfirm the causal power. White argues then that the proportion of *a*+*d* cells from among the entire set (*a, b, c,* and *d*) serves to modulate participants' judgments. When there is a higher PCI, judgments will increase. Faced with a relatively low PCI, judgments will be attenuated. This approach also seems inherently compatible with the concepts of confidence and reliability. Both reliability and PCI formally express how a focal sub-set of the trials deserve more attention from the causal reasoners. It follows that when either reliability or PCI is elevated, confidence in the judgments will also increase.

In looking at goodness-of-fit, ΔP and the R-W model will also be reconsidered from this different perspective. Each alternative under consideration will also be combined with reliability, to observe the effect of merging the concepts together. Below, Table 8 presents the bivariate correlation coefficients for each of the approaches to the

obtained judgments. First, these correlations will be considered for each experiment separately, then across all four to see if there is one approach which best fits all of the experimental data regardless of between-experiment variances. Note that to obtain the predictions of the R-W model, ß-values were calculated according to the least squared differences approach between the obtained and R-W predictions. This resulted in a ß-effect present weight of .72 and a ß-effect absent weight of .76. There was no a-priori rationale for these values; however they represent the best fit with the actual judgments. These values are also particularly useful because they are extremely close, and therefore would appear to meet the requirements of the more conservative *restricted*-R-W model. The assumption of equal ß-weights is inherent in the restricted model, and is also an assumption of the Power PC theory.

Table 8: Assessing the Fit of Models to the Obtained Judgments

Model	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Overall
PCM	-.06	.995**	.173	.913**	.702*
Power PC	.935**	N/A [1]	.774**	.932**	.864**
PCI	.948**	-.995**	.174	.584*	.401*
R-W	.031	.995**	.182	.913**	.707**
Reliability	-.916**	.995**	-.042	.007	.076
PCM*Reliability	-.679*	.988**	.096	.860**	.596**
PCI*Reliability	.955**	-.547	.183	.506	.350*
R-W*Reliability	-.672*	.989**	.097	.860**	.597**

* = significant at .05 level, ** = significant at .01 level.,
[1] Power PC held constant in Experiment 2

This analysis of the relative fit of each model's predictions to the obtained judgments reveals a number of trends. Please note again that these are post-hoc analyses, and only the PCM and Power PC were controlled for systematically across all four

experiments. Experiment 1 fits best with the PCI model's predictions. Combining the concept of reliability with the PCI model yields even more impressive results. Recall that this experiment produced a pattern very consistent with the Power PC model. This analysis further supports that finding. The second experiment set out to determine whether the Power PC predictions were sensitive to the actual judgment levels, rather than just the pattern. Power was held constant for this experiment, yet the majority of the other models are sensitive to the judgment variations produced by the participants. Interestingly, the PCI predictions are almost completely opposite to the obtained values.

The third experiment was in some ways the inverse of Experiment 2, as it provided maximal contrast between predicted power levels, while controlling for reliability and overall probability of the cause. Here, the Power PC model provides the strongest correlation to the obtained data. This post-hoc analysis does not pick up on the interactions between the experimental factors however. And again, as shown through Experiments 1 and 2, the Power PC model on its own cannot account for the actual obtained judgment levels. Recall again the demonstration that Power PC incorporating Reliability ultimately reduces to the PCM.

Experiment 4 shows several strong competing candidates. This experiment was designed to show how the discoveries from Experiment 1 and 2 would hold up in negative contingency space. The Power PC theory again shows a high degree of correlation, yet both the PCM and associative R-W models are also highly correlated to the judgments. The correlation levels for this experiment are quite similar to the correlation levels overall. The Power PC theory continues to be statistically relevant, yet the weaknesses as demonstrated in the first two experiments remain valid criticisms. The

addition of the concept of reliability is intended to account for those conditions which the Power PC theory cannot account for on its own, yet that particular combination results in a return to the PCM. For the most part, incorporating reliability with pre-existing models tends to fall short, with the exception of the combination of PCI and reliability in Experiment 1.

Experiment 3 demonstrated a strong link between reliability and confidence, suggesting that future research ought to incorporate reliability in the experimental design, and collect confidence through the experimental task. This will provide for a deeper understanding of the interaction of these factors, and indeed whether there is a particular directional causal relation between the two of them. The post-hoc analysis presented here seems to indicate that reliability does not fit best with any particular current model, and that the opportunity to incorporate reliability into a new model, possibly involving confidence, may prove to be the most fruitful.

Chapter 4: Conclusion

This report presented the results of four original experiments aimed at exploring the factors which affect human causality judgment. Prior research had demonstrated that in some circumstances it is difficult to distinguish the predictions of associative and computational models. The Power PC theory has risen to prominence in recent years, yet the evidence weighs both for and against its adoption a golden standard. Confidence and reliability are two factors which have been scarcely documented in the literature, yet may provide key insights about the causal attribution process.

This series of experiments has demonstrated that the Power PC theory on its own, or in conjunction with reliability, is an insufficient model of human contingency judgment. While the Power PC theory may account for some particular results that other computational and associative models cannot, it falls short of being a complete account of causal attribution, for both generative and preventive candidate causes. Beyond this discovery, there are avenues for future research which show promise. The data presented here on confidence, on its own and in relation to reliability, is of particular interest and it would be valuable to see other researchers begin to incorporate these concepts into the design and analysis of their experiments. There were some relatively anomalous results from Experiment 4 (dramatic drop in judgments at $\Delta P = .3$) which seems to point to a potential role for minimal occurrence of the effect in the absence of the cause. The conditions demonstrating this drop all involved zero cases where the effect occurred in the absence of the cause.

The conceptual framework for most current theories of human contingency judgment has been either associative or computational. With the Power PC theory shown

as inadequate, future knowledge in this area may be gained from bridging the two approaches, or moving outside the current boundaries. An ideal theory ought to be able to account for patterns and levels of judgment, across both positive and negative contingencies. Additionally, such a theory should work equally well for both categorical and continuous causes and effects. Currently, no single theory can account for human contingency judgments in all of these conditions.

In recent years, several alternate views of the causal attribution process have been proposed. Several of these seek to transcend the inherent limitations of the traditional R-W, PCM or Power PC theories here. Bayesian nets (Danks, 2005; Griffiths & Tenenbaum, 2005), are proposed as a potential resolution to the longstanding argument between associative (R-W model) and computational (PCM, Power PC) approaches to the causal attribution task. In particular, Danks (2005) argues that a single solution might not be feasible, as different causal judgments may rely on different fundamental processes and require different normative models. Indeed, this is how Griffiths and Tenenbaum (2005) have characterized their causal support model. While traditional approaches such as R-W, PCM, and Power PC all act as normative models for the question "How strong is the causal relation between the candidate cause and effect?", causal support answers the more basic question of "Is there a causal relation between the candidate cause and effect?". For this reason, while the causal support model appears promising, it cannot be taken as a replacement for earlier approaches since it addresses a separate cognitive process. If future studies incorporate both the task of determining whether there is a causal relation and the strength of that relation, it will be possible to

learn how the two processes interact. Doing so may provide the missing evidence in our current understanding of the causal attribution process.

Another approach discussed in the previous section, but which has received relatively little experimental attention, is the Proportion of Confirming Instances (PCI) approach of White (1998, 2000, 2001, 2003, 2004). The PCI rule essentially results in an alternate-weighting of the PCM model, on the basis that for generative causes, cells *a* and *d* represent 'confirmation' of the generative power of the cause, while cells *b* and *c* serve to disconfirm the causal power.

Signal detection theory (SDT) has only recently been applied in this context, but shows promise as well (Allan, Siegel & Tangen, 2005). This involves borrowing a theory which has normally been applied primarily to psychophysics questions, and incorporating it into the binary contingency judgment task. The authors demonstrate how a novel SDT approach can account for the prevalent outcome-density bias in causal attribution. Recall that this effect relates to increasing judgments as the overall probability of the effect increases. It is argued that this demonstrates how the SDT can transcend its origins and find meaningful applications in the realm of causality, similar to how the R-W model transcended the Pavlovian animal-learning roots to explain human causal attribution and decision-making. Much like the causal support approach, the SDT is not a perfect fit for the current methodological paradigm but with some modifications, future studies may be better suited to evaluate its usefulness.

At this point, all of these alternatives are lacking widespread empirical support. Among them, researchers may find an integrated solution to causal attribution. It also seems that any successful model will need to resolve the currently undetermined roles of

both confidence and reliability. In light of the evidence presented here, both constructs deserve to be taken into account for their apparent role in the modulating contingency judgments.

References

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. (2003). Assessing power pc. *Learning & Behavior, 31*(2), 192-204.

Allan, L. G.; Seigel, S. & Tangen, J. (2005). A signal-detection analysis of contingency data. *Learning and Behaviour, 33*(2), 250-263.

Alloy, L. B. & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91*, 112-149.

Anderson, J. R. & Sheu, C. (1995). Causal inferences as perceptual judgments. *Memory and Cognition, 23*(4), 510-524.

Baker, A. G. Berbrier, M. & Vallée-Tourangeau, F. (1989). Judgments of a 2x2 contingency table: Sequential processing and the learning curve. *Quarterly Journal of Experimental Psychology, 41*B, 65-97.

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments - presence of a strong causal factor may reduce judgments of a weaker one. *Journal Of Experimental Psychology-Learning Memory And Cognition, 19*(2), 414-432.

Baker, A. G., Murphy, R. A., & Vallée-Tourangeau, F. (1996). Associative and normative models of causal induction: Reacting to versus understanding cause. In *Causal learning* (Vol. 34, pp. 1-45).

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

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.

Cheng, P. W. & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and Luhmann and Ahn (2005). *Psychological Review, 12*(5), 694-707.

Cheng, P. (1992). Covariation in natural causal induction. *International Journal Of Psychology, 27*(3-4), 173-173.

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

Cheng, P. W. (1993). Separating causal laws from casual facts - pressing the limits of statistical relevance. In *Psychology of learning and motivation: Advances in research and theory, 30*, pp. 215-264.

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

Clément, M., Mercier, P., & Pastò, L. (2002). Sample size, confidence, and contingency judgement. *Canadian Journal of Experimental Psychology, 56*(2), 128-137.

Danks, D. (2005). The supposed competition between competing theories of human causal inference. *Philosophical Psychology, 18*(2), 259-272.

Dennis, M. J., & Ahn, W. K. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition, 29*(1), 152-164.

Fugelsang, J. A., & Thompson, V. A. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology, 54*(1), 15-32.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and support in causal induction. *Cognitive Psychology, 51*, 334-384.

Horne, M. (2002). Aristotle's ontogenesis: A theory of individuation which integrates the classical and developmental perspectives. *Journal of Analytical Psychology, 47*(4), pp. 613-628

Horne, M.; Sowa, A.; & Isenman, D. (2000). Philosophical assumptions in Freud, Jung and Bion: Questions of causality. *Journal of Analytical Psychology, 45*(1), 109-121.

Iturrate, M. (1976). Man as the meaning-giver: An existential-psychoanalytic approach. *Review of Existential Psychology and Psychiatry, 14*(2), 63-80.

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

Kao, S. F., & Wasserman, E. A. (1996). Information integration and associative learning as accounts of human contingency judgment. *International Journal Of Psychology, 31*(3-4), 1411-1411.

Kohlberg, L. (1968). The child as a moral philosopher. *Psychology Today, 2*(4), 25-30.

Lagnado, D. A., & Shanks, D. R. (2002). Probability judgment in hierarchical learning: A conflict between predictiveness and coherence. *Cognition, 83*(1), 81-112.

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

Lopez, F. J., Shanks, D. R., Almaraz, J., & Fernandez, P. (1998). Effects of trial order on contingency judgments: A comparison of associative and probabilistic contrast accounts. *Journal Of Experimental Psychology-Learning Memory And Cognition, 24*(3), 672-694.

Lopez, G. F. J., Fernandez, P., & Lopez, G. (1992). The role of the task in contingency judgment. *International Journal Of Psychology, 27*(3-4), 123-123.

Luhmann, C. & Ahn, W. (2005). The meaning and computation of causal power: Comment on Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 12*(5), 685-693.

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

Mandelbrot, B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science, 156*(3775), 636-638.

Matute, H., Vegas, S., & De Marez, P. J. (2002). Flexible use of recent information in causal and predictive judgments. *Journal Of Experimental Psychology-Learning Memory And Cognition, 28*(4), 714-725.

Mercier, P. (1996). Computer simulations of the Rescorla-Wagner and Pearce-Hall models in conditioning and contingency judgment. *Behavior Research Methods Instruments & Computers, 28*(1), 55-60.

Mercier, P., & Parr, W. (1996). Inter-trial interval, stimulus duration and number of trials in contingency judgments. *British Journal Of Psychology, 87*, 549-566.

Natural Resources Canada (2000). *Coastline by Province and Territory.* Retrieved December 6, 2005, from http://atlas.gc.ca/site/english/learningresources/facts/coastline.html

Nissim-Sabat, M. (1999). Phenomenology and mental disorders: Heidegger or Husserl? *Philosophy, Psychiatry, and Psychology, 6*(2), 100-104.

Novick, L. R., Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*(2), 455-485.

Over, D. E., & Green, D. W. (2001). Contingency, causation, and adaptive inference. *Psychological Review, 108*(3), 682-684.

Pacherie, E. (2004). Looking for the agent in action. *Trends in Cognitive Sciences, 8*(2), pp. 54-55.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review Of Psychology, 52*, 111-139.

Phillips, A. T. (2004). The ontogenesis of the understanding of intentional and physical causality. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 64*(8-B), 4079.

Piaget, J. (2001, orig. 1930). *The Child's Conception of Physical Causality*. Transaction Publishers, New Brunswick, NJ, USA.

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.

Romdhane, L. B., Ayeb, B. (2005). A distributed neural approach for causal reasoning using cooperative and competitive neural computations. *Applied Artificial Intelligence, 19*(5), pp. 535-557

Shanks, D. R. (1985a). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition, 13*(2), 158-167.

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

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning And Motivation, 18*(2), 147-166.

Shanks, D. R. (1991). On similarities between causal judgments in experienced and described situations. *Psychological Science, 2*(5), 341-350.

Shanks, D. R. (1995). Is human learning rational. *Quarterly Journal Of Experimental Psychology Section A-Human Experimental Psychology, 48*(2), 257-279.

Shanks, D. R. (2002). Tests of the power pc theory of causal induction with negative contingencies. *Experimental Psychology, 49*(2), 81-88.

Shanks, D. R., Charles, D., Darby, R. J., & Azmi, A. (1998). Configural processes in human associative learning. *Journal Of Experimental Psychology-Learning Memory And Cognition, 24*(6), 1353-1378.

Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In *Causal learning* (Vol. 34, pp. 265-311).

Shimazaki, T. (1999). On the validity of applying associative learning model to the acquisition process of human contingency judgment. *Japanese Journal Of Psychology, 70*(5), 409-416.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677-680.

Thoma, H. & Cheshire, N. (1991). Freud's nachtraglichkeit and Strachey's 'deferred action': Trauma, constructions and the direction of causality. *International Review of Psycho-Analysis, 18*(3), 407-427.

Vallée-Tourangeau, F., Hollingsworth, L., & Murphy, R. A. (1998). 'Attentional bias' in correlation judgments? Smedslund (1963) revisited. *Scandinavian Journal Of Psychology, 39*(4), 221-233.

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.

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. *Quarterly Journal Of Experimental Psychology Section A-Human Experimental Psychology, 51*(1), 65-84.

Vallée-Tourangeau, F. R., Murphy, R. A., & Baker, A. G. (1998). Causal induction in the presence of a perfect negative cue: Contrasting predictions from associative and statistical models. *Quarterly Journal Of Experimental Psychology Section B-Comparative And Physiological Psychology, 51*(2), 173-191.

Vallée-Tourangeau, F., Baker, A. G., & Mercier, P. (1994). Discounting in causality and covariation judgments. *Quarterly Journal Of Experimental Psychology Section B-Comparative And Physiological Psychology, 47*(2), 151-171.

van Osselaer, S. M. J., Janiszewski, C., & Cunha, M. (2004). Stimulus generalization in two associative learning processes. *Journal Of Experimental Psychology-Learning Memory And Cognition, 30*(3), 626-638.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82*(1), 27-58.

Wasserman, E. A. (1996). Contingency judgment: Data, theory, and implications for psychological science. *International Journal Of Psychology, 31*(3-4), 1410-1410.

Wasserman, E. A., Kao, S. F., VanHamme, L. J., Katagiri, M., & Young, M. E. (1996). Causation and association. In *Causal learning* (Vol. 34, pp. 207-264).

Wasserman, E. A., & Miller, R. R. (1997). What's elementary about associative learning? *Annual Review Of Psychology, 48*, 573-607.

White, P. A. (1998). Causal judgment: Use of different types of contingency information as confirmatory and disconfirmatory. *European Journal Of Cognitive Psychology, 10*(2), 131-170.

White, P. A. (2000). Causal judgment from contingency information: Relation between subjective reports and individual tendencies in judgment. *Memory & Cognition, 28*(3), 415-426.

White, P. A. (2001). Causal judgments about relations between multilevel variables. *Journal Of Experimental Psychology-Learning Memory And Cognition, 27*(2), 499-511.

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*(4), 710-727.

White, P. A. (2004). Causal judgment from contingency information: A systematic test of the pci rule. *Memory & Cognition, 32*(3), 353-368.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science, 10*(2), 92-97.

Yamaguchi, M. (1999). New methods for solving the Rescorla-Wagner model. *Behavior Research Methods Instruments & Computers, 31*(4), 684-688.

Young, M. E. (1995). On the origin of personal causal theories. *Psychonomic Bulletin & Review, 2*(1), 83-104.

Zorzi, M. (2004). Routes to reading success and failure: Toward an integrated cognitive psychology of atypical reading. *European Journal of Cognitive Psychology, 16*(3), 475-477.

APPENDIX A

University of Ottawa Research Ethics Approval

NOTE TO USERS

APPENDIX B

Research Participant Consent Form

CONSENT FORM

RESEARCHERS: Pierre Mercier, PhD (562-5800, ext. 4292)
Matt Jacques (562-5800, ext. 4289)

By respect for the rights of individuals, the University of Ottawa and the research granting agencies require the written consent of the research participants.

This research project is carried out by Matt Jacques, doctoral student, under the supervision of Pierre Mercier, Ph.D. The project entails no physical or psychological risks other than the visual fatigue that may be associated with watching a computer screen for 45 minutes. The judgments that we ask you to make might sometimes appear difficult but we are only interested that you do your best. We also ask that you make the mental effort necessary in using the full capacity of your memory.

I, _____, am consenting to participate in this study on judgement and decision making carried out under the supervision of Pierre Mercier, Ph.D and Matt Jacques. The main objective of this study is to further our understanding of judgement processes.

My participation will consist in attending an individual session of about 45 minutes at the Cognitive Psychology Laboratory in which I will have to evaluate the relation between two symbolic events [symptom and disease] presented many times on a computer screen. I agree to let the data be used for research purposes as long as I am not identified personally in the report. My name appears only on this consent form.

I am free to withdraw from the study at any time without any form of penalty whatsoever. Any question concerning my rights as a research participant can be addressed by contacting Catherine Lesage, Protocol Officer for Ethics in Research at (613) 562-5800 ext. 1787 or clesage@uottawa.ca.

SIGNATURES

PARTICIPANT:_____ DATE:_____

RESEARCHER:_____ DATE:_____

APPENDIX C

Symbols and Abbreviations Used

Symbol or Abbreviation	Definition or Explanation
1. ΔP	Delta-P, the basic measure of correlation between a cause and effect when expressed in a 2x2 contingency table. Serves as a normative parameter for human contingency judgment under the PCM model, and is therefore often referred to in this text as "contingency".
2. PCM	Probabilistic Contrast Model of human contingency judgment. Falls under the category of computational model, as it is based on the assumption that causal reasoners compute ΔP to arrive at their decisions.
3. R-W or RWM	The Rescorla-Wagner model, an influential associative model of contingency judgment.
4. p_i	Causal power, the normative parameter of causal attribution prescribed by the Power PC model.
5. Power PC	The Power version of the PCM, shortened as Power PC. This model is intended to correct several of the shortcomings of other computational and associative accounts of the contingency judgment process.
6. R	Reliability, the number or proportion of trials when the cause is present and the effect occurs, which cannot be explained by the overall base rate of the effect occurring in the absence of the cause. There may be any number of alternate causes which can bring about the effect regardless of whether the candidate cause is present. Reliability is intended to control for or rule out trials on which the effect could be said to be due to alternate causes.