National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre reference

Our file Notre reference

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Canada

COVARIATION DETECTION BIASES IN SUFFICIENT AND NECESSARY SITUATIONS

by

Yuanshan Cheng
School of Psychology
University of Ottawa

A doctoral dissertation submitted to
the School of Graduate Studies of the University of Ottawa

In partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Experimental Psychology

August, 1995

ISBN 0-612-07842-6

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Acknowledgements

In the opportunity of presenting my dissertation, I would like to express my deepest appreciation to Dr. Pierre Mercier. As a foreign student commencing a doctoral program, I was in need of various forms of support. Dr. Mercier provided me with invaluable assistance in language improvement, cultural adjustment, course selection, and financial support. At all stages in fulfilling the requirements of the program, he also helped me to learn various programming languages, to develop my thesis topic, design the necessary experiments, and perform the appropriate statistical analyses, and to polish my writing style. I strongly feel that without his help, I would not have finished my program.

I would also like to thank the members of my thesis committee for their constructive feedback and suggestions during all phases of completing my dissertation.

COVARIATION DETECTION BIASES IN SUFFICIENT AND NECESSARY SITUATIONS

Yuanshan Cheng

Abstract

In 4 experiments, university students played video games in which one action or cause covaried with an outcome. Judgments on sufficient and necessary causes were observed. On the basis of the obtained judgments, different computational models, Cheng and Novick's (1990a, 1992) probabilistic contrast (ΔP rule) and the Rescorla-Wagner (1972) model were evaluated. In Experiments 1 and 2, for the positive contingencies, the participants judged sufficient and necessary causes differently; they also showed judgment deviations from the real contingencies. The ΔP rule could not account for these data. An alternative weighted ΔP rule was proposed and, along with the Rescorla-Wagner model, it successfully explained these results. In Experiment 3, negative contingencies were included. The pattern of judgments among the negative sufficient and necessary causes mirrored that of the positive contingencies but did not reach statistical significance. The ΔP rule could not account for the judgments in Experiment 3, the adjusted ΔP rule did not either. However, the Rescorla-Wagner model accounted for the results very well. In Experiment 4, the predictive power of these different models was compared. In general, the Rescorla-Wagner model remains the best descriptive model for explaining and predicting the patterns of contingency judgments.

Table of Contents

List of Figures

List of Tables

COVARIATION DETECTION BIASES IN SUFFICIENT AND NECESSARY SITUATIONS

Introduction

An outcome may result from one or several possible causes or actions. In daily life, people predict the occurrence of one event from the others when these events covary. People are also sensitive to the fact other events can be independent of one another. This ability to detect contingent and noncontingent relations has allowed human beings to survive and succeed in the world. Even very young children may already have some experience at detecting these covariations. For instance, Shultz and Mendelson (1975) found that children as young as 3 years of age were capable of using covariation information in their attributions of simple physical effects. Thus, the ability to detect covariations among events appears to develop very early and to constitute an important adaptive skill. And, considering the extent of the literature on animal conditioning (see Dickinson, 1980; Mackintosh, 1974), the same can be said for many, if not all, animal species.

Covariation Detection in Different Subareas of Psychology

Scientists from different areas of psychology have suggested that judgment of covariation is a crucial cognitive task (Alloy & Abramson, 1979; Seligman, 1975; Shaklee & Mims, 1982; 1986; Shaklee & Tucker, 1980), and some consider it as part of the larger topic of problem solving (Anderson, 1990a; 1990b; Newell, 1980). Recently, there has been considerable interest in the cognitive underpinnings of human covariation detection (e.g., Baker, Mercier, Vallée-

Tourangeau, Frank, & Pan, 1993; Dickinson, Shanks, & Evenden, 1984; Cheng & Novick, 1990a; 1992; Gluck & Bower, 1988; Kahneman & Tversky, 1973; Kao & Wasserman, 1993; Shanks 1985a; 1985b; 1986; 1991; Tversky & Kahneman, 1974; Wasserman, Elek, Chatlosh, & Baker, 1993).

Pursing a different objective, developmental psychologists posited that as children grow up, they develop the use of rules in their judgment of covariations (e.g., Inhelder & Piaget, 1958; Shaklee & Mims, 1982, 1986; Shaklee & Tucker, 1980). Some studies connected with clinical psychology reported that some difficulties of adaptation, such as depression, might influence or be influenced by covariation detection (Alloy & Abramson, 1979; Seligman, 1975; Tang & Critelli, 1990). In the area of social psychology, there are a large number of studies addressing causal judgments as social attributions (see Cheng & Novick, 1990a; 1990b; 1992; Hewstone, 1983; Kelley, 1972). The attribution literature focuses on the conditions under which people attribute outcomes to different kinds of causes (e.g., internal vs. external, self vs. others, situation vs. dispositions). The relevant theories concern the different causal dimensions of the covariation situation. For instance, Kelley (1967, 1973) proposed an ANOVA model. He considered that people were "intuitive scientists" who used a mechanism of causal induction analogous to the analysis of variance. He also suggested that in explaining human social attribution, the relevant causal dimensions were persons (P), stimuli (S), and times (T). Kelley described these dimensions as the independent variables in his ANOVA model, and proposed that one infers the cause of a given P's response to a certain S on a particular occasion T depending on one's perception of the degree of: a) *consensus* between P's response to S and other people's responses to S, b) *distinctiveness* of P's response to S from P's responses to other stimuli, and c) *consistency* of P's

response to S at other times.

Measures of Covariation and the ΔP Rule

In the area of cognitive psychology, human judgments of covariation have been involved in studies of judgment under uncertainty, causal inference, estimation of relatedness, probability judgments, and contingency judgments. The tasks used in these studies usually involve covarying dichotomous variables, and the simplest one includes only two such dichotomous variables. For example, one of the variables can be an *action* and the other is the *outcome* of the action. This is also called a *single contingency task*. The relationship between these two variables (outcome and action) can be represented by a 2 x 2 contingency table (see Table 1). Such a table clearly shows the frequencies of the presence and absence of an outcome when an action, or "cause", is given or not given. The letters A, B, C, and D in the contingency table represent the frequencies of the four possible action-outcome combinations respectively.

Table 1

A 2 X 2 contingency table for an action-outcome contingency. The letters A-D represent the cumulative frequencies of the two variable combinations.

	OUTCOME (O)	NO OUTCOME (~O)
WITH ACTION (A)	**A**	**B**
WITHOUT ACTION (~A)	**C**	**D**

A number of normative measures have been proposed to quantitatively compute the

covariation between two dichotomous variables. For instance, Smedslund (1963) referred to the covariation as Δr_1 and Δr_2, which are calculated by employing the ratio of the four cell frequencies:

$$\Delta r_1 = (A+D)/(B+C) \quad (1)$$

$$\Delta r_2 = (A+D)/(A+B+C+D) \quad (2)$$

Instead of using the ratio of cell frequencies, Inhelder & Piaget (1958) suggested that the sum of diagonals, the ΔD rule, was a reasonable measure of the covariation:

$$\Delta D = (A+D) - (B+C) \quad (3)$$

Allan (1980) has convincingly argued that the most appropriate measure of the covariation of two binary variables is ΔP coefficient, which was proposed by Jenkins and Ward in 1965. This coefficient is the only one that can capture relationships based on all possible combinations of frequencies in the 2 X 2 table. It is the difference between two conditional probabilities:

$$\Delta P = P(outcome \mid action) - P(outcome \mid without\ action)$$

$$= A/(A+B) - C/(C+D) \quad (4)$$

In recent studies of covariation detection, ΔP is generally used as the objective criterion to evaluate the accuracy of subjective judgments at different levels (Allan & Jenkins, 1980; 1983; Baker, Berbrier, & Vallée-Tourangeau, 1989; Baker & Mercier, 1989; Baker, et al., 1993;

Dickinson, et al., 1984; Shanks 1985a; 1985b; 1986; 1987; Kao & Wasserman, 1993; Shanks & Dickinson, 1987; Vallée-Tourangeau, Baker, & Mercier, 1994; Wasserman, Chatlosh, & Neunaber, 1983; Wasserman, et al., 1993). In these studies, subjects' judgments were considered biased when they differed from the target ΔP values.

Methodological Issues in the Study of Contingency Judgments

1). Group data and individual data

In studying contingency judgments, researcher must make two importantly methodological decisions. The first is whether to look at individual data or to focus on group data. The second is whether to use a paradigm in which the participants experience all the trials directly, hereafter called "on-line", or to employ a paradigm in which the participants are exposed to pieces of information summarized in advance for them, hereafter called "off-line".

Some researchers have chosen to study contingency judgments at a *group level* (e.g. Baker, et al., 1993; Dickinson, et al., 1984; Shanks 1985a; 1985b; 1986; 1987; Kao & Wasserman, 1993; Shanks & Dickinson, 1987; Vallée-Tourangeau, et al., 1994; Wasserman, et al., 1993). In these studies, researchers use ΔP values as a criterion to evaluate the mean judgments of a group of subjects. With the help of statistical analyses, they then try to infer the judgment process at the level of population means.

Another category of the studies focuses on the performance of individuals. These studies assume that different individuals may use different judgment rules and attempt to classify people as different rule users. For example, Shaklee and Mims (1982) and Shaklee and Tucker (1980) investigated four possible judgment rules: cell A, cell A versus cell B, ΔD, and ΔP. According

to the cell A rule, users judge the relationship between two variables based on the frequency of cell A. The "cell A versus cell B" rule (or its variant "cell A versus cell C" rule) users make their judgments based on the difference between the respective frequencies of these cells (either A - B or A - C). These rules are not considered normative. According to Inhelder and Piaget (1958) they are developmental precursors to the more mature ΔD rule, which itself is superseded by the ΔP rule and considered as non-normative too.

To test use of these possible rules, Shaklee and Tucker (1980) showed sets of problems to their subjects. One set of stories involved people being healthy or sick as a function of the possible presence or absence of an injection, liquid medicine, or a pill. These problems were organized in 2 X 2 tables, as illustrated in Table 2. According to the rule analytic technique, if a subject uses the cell A rule, he or she can only judge the relationship correctly on the cell A problems. If a subject uses the cell A versus B rule, he or she can give correct answers to the cell A and the cell A versus B problems. For the ΔD rule users, all problems, except of the ΔP problems, can be judged correctly. Only the ΔP rule users can give correct answers to all problems. This kind of approach has been repeatedly employed on the basic assumption that, in the same experiment, each individual subject consistently uses a single rule to make his or her judgments.

Table 2

Cell frequencies for the problems used by Shaklee and Tucker (1980). The ΔP values are positive for the four tables in the top row, null for the middle row, and negative for the bottom row.

Cell A problems			Cell A vs. B problems			ΔD problems			ΔP problems		
	B1	B2		B1	B2		B1	B2		B1	B2
A1	11	4	A1	4	1	A1	4	4	A1	2	12
A2	1	8	A2	3	16	A2	1	15	A2	0	10
	B1	B2		B1	B2		B1	B2		B1	B2
A1	6	6	A1	4	4	A1	9	5	A1	1	5
A2	6	6	A2	8	8	A2	7	3	A2	3	15
	B1	B2		B1	B2		B1	B2		B1	B2
A1	1	8	A1	4	11	A1	4	4	A1	12	2
A2	11	4	A2	8	1	A2	15	1	A2	10	0

Based on this kind of study, Inhelder and Piaget (1958) suggested that the cell A rule is the strategy used by younger adolescents (12-13 years), whereas older adolescents (14-15 years) would rely on the ΔD rule. Other studies showed that adults may use the ΔD or the ΔP rules. Nevertheless, the cell A rule often remains a typical strategy used by the same adolescents and adults (e.g., Shaklee & Tucker, 1980; Smedslund, 1963). These experiments seem to reveal that most individuals use one of those postulated rules to make their judgments, and different individuals may use different rules. However, this conclusion is incompletely warranted, since at least, the measurement used in these experiments is only qualitative. These studies are not

sufficient to infer people's judgment processes over a wide range of subtle quantitative covariations.

Additionally, as attractive as the rule analytic technique appears, these studies determine rule use in a post hoc, empirical manner. While they purport to show that different individuals use different rules and even that the same individual may change rules over time, they cannot specifically predict when a given rule is going to be used. It is possible that some or all of the individuals use another rule or computational procedure, which has not been identified but can singly account for all the judgments. There is strong support for this criticism in the fact that a sizeable proportion of subjects cannot be categorized as using any of the four rules tested by Shaklee and her collaborators (Shimazaki, Tsuda, & Imada; 1991).

One possible computational procedure incorporating all the normative rules mentioned above was reported by Kao and Wasserman (1993). They created a series of noncontingency problems and asked their subjects to make quantitative judgments in the experiments. In analyzing their data, they applied a "parameter search" technique, called the Hooke and Jeeves method (HJ method) (see Kao & Wasserman, 1993). This method systematically searches all parameter combinations to find the best set of weights for a specific formula adapted from the original work of Busemeyer (1991). The equation is shown as follows:

$$R = (W_A \times A - W_B \times B - W_C \times C + W_D \times D) / (W_A \times A + W_B \times B + W_C \times C + W_D \times D) \quad (5)$$

In this equation, A, B, C, and D denote the frequencies of the four cells in the contingency table;

Ws denote the weights for different cells. The sum of the four weights equals 1. There is a close resemblance between this equation and the ΔD rule. The numerator is the same as ΔD except that each cell is multiplied by a weight, and the denominator of the formula can be treated as a proportion for the overall ΔD value. Kao and Wasserman (1993) showed that the equation was the best formula to differentiate individual subjects into different rule users, since by setting different weights to different cells, it was possible to reproduce all the non-normative rules. For instance, if the weights for cells B, C, and D are very small, these cells can be ignored in the equation and only cell A remains in the formula; thus, this formula becomes the cell A rule. On the other hand, if the weights for cells C and D, or for cells B and D, are very small, the cell A versus cell B rule or the cell A versus cell C rule is obtained. If all four weights are almost equal to .25, the ΔD rule obtained.

Using the HJ method, Kao and Wasserman (1993) calculated the best weights for each research participant individually and, with a set of practical criteria classify them as different rule users. For example, the criteria $W_A \geq .70$ and $W_B, W_C, W_D \leq .10$ classified a subject as a cell A rule user; $.40 \leq W_A, W_B \leq .60$, and $W_C, W_D \leq .10$ classified a subject as a user of the cell A versus cell B rule. Even under these practical criteria, only 36% of their subjects could be classified as those postulated rule users. These results were consistent with other similar studies (Allan & Jenkins, 1980; Arkes & Harkness, 1983; Capon & Kuhn, 1979; Shaklee & Hall, 1983; Shaklee & Mims, 1981; 1982; Shaklee & Tucker, 1980; Ward & Jenkins, 1965).

One seemingly straightforward method to identify people's judgment strategies might have been to simply ask them how they make their judgments. The validity of this method rests on the assumption that subjects are able and willing to accurately describe their judgment processes.

However, several studies indicate that this assumption is unjustified: the subjects describe themselves as using complex rules that bear little resemblance to their simple judgment patterns (e.g., Adi, Karplus, Lawson, & Pulos, 1978; Goldberg, 1968; Ericsson & Simon, 1980).

As discussed above, the entire rule analytic paradigm is of questionably informational value because even the most sophisticated and individualized categorization technique leaves too many participants unclassified. The paradigm over-emphasizes individual differences to the detriment of a successful search for a general computational mechanism. Without negating individual differences, the experiments presented below return the focus to the search for a general mechanism.

2). Off-line and on-line studies

Another important methodological feature of contingency judgment studies is that many of them were conducted by an "off-line" paradigm, in which the subjects are not required to directly observe the covariations of the outcome and the action(s) as they occur. Rather, they are told one or a few stories in which the outcome and the action(s) are presented as daily life events. The frequencies of the two variable combinations are usually either described in the stories, or directly shown in tabulated form (e.g., Kao & Wasserman, 1993; Shaklee & Mims, 1982; Shaklee & Tucker, 1980; Ward and Jenkins, 1965). These off-line studies also suffer from a lack of ecological validity. Even using daily life events as the actions and the outcomes, off-line studies are still fairly different from real life situation. In these experiments the combinations of the two variables are not ◦ perienced in an ongoing fashion as they generally are on a day to day basis. Books, news items, or scientific reports are more likely to present summarized or

tabulated information. Yet the ability to detect covariations which has evolved naturally does not normally rely on tabulated input. Indeed, it has been observed, both across (Arkes & Harkness, 1983; Arkes & Rothbart, 1985; Jennings, Amabile & Ross, 1982; Shaklee & Mims, 1986) and within experiments (Kao & Wasserman, 1993; Schustack, 1988; Ward & Jenkins, 1965), that off-line judgments are more consistent with ΔP than on-line judgments. This finding suggests that different processes might be induced by the different tasks.

Because there is a distinct possibility that off-line studies do not tap into the natural covariation detection processes and that on-line research might be more revealing in this regard, the experiments reported in this thesis use the on-line method.

In most, though not all, on-line experiments, the subjects are asked to produce the action themselves in order to keep their attention, and then, to observe whether the outcome occurs or not. After observing a set of trials, they are usually required to estimate the strength of the covariation between the two variables on a scale. Thus, their judgments are not only relevant to qualitative relations, but also to quantitative ones. This procedure appears ecologically more valid since the events must be processed as they are experienced.

Several on-line studies (Alloy & Abramson, 1979; Allan & Jenkins, 1980, 1983; Baker et. al., 1989; Dickinson et. al., 1984; Shanks, 1987; Wasserman et. al., 1983, 1993) have demonstrated that, in general, human judgments of covariations are fairly close to the normative contingencies as calculated by the ΔP rule, although, sometimes, some kinds of judgment biases do exist. For example, by using a game like computer program, Dickinson and his colleagues (1984) found that changes in the subjects' judgments were consistent with the ΔP value. In the games, on some trials, when a tank passed across the screen, the subjects were asked to press a

key on the keyboard to fire a shell at the tank. In this game then, firing the shell was an action, and also one possible cause of the tank's explosion. There was also a minefield on the screen, which the tank had to pass through. The mines in the field might also cause the explosion of the tank. Thus, without firing, the tank might also explode. The mines may be referred to as another cause of the outcome, or if we consider the situation without the firing of shells as the context, the mines may be referred to as a contextual cause. Note that, in the games, if a tank was to explode, that would happen in the minefield; thus from the explosion itself the subjects could not tell if it was caused by the shell or by the mines. Six different games, each including 40 trials, were employed in this experiment. The ratio of P(O|A), the probability for the outcome given the action, and P(O|~A), the probability of the outcome without the action, was respectively .75 : .25, .75 : .50, .75 : .75, .25 : .25, .50 : .75, and .25 : .75. Thus, the corresponding ΔP values for the six games were .50, .25, 0, 0, -.25, and -.50. After every 40 trials, the subjects were required to judge the effectiveness of the shell by using a -100 to +100 scale. The results generally showed that the subjects' judgments were consistent with ΔP values: when P(O|A) was held constant, subjects' judgments decreased as P(O|~A) increased; conversely, when P(O|~A) was kept constant, subjects' judgments decreased as P(O|A) also decreased. For the non-contingency conditions, however, the subjects' judgments deviated from zero. These judgment biases were linked to the frequency of the outcome and called "outcome density bias". When the frequency of the outcome was high (.75 : .75), a positive bias appeared: subjects judged the zero contingency as positive. However, when it was low (.25 : .25), the judgment became negative.

In the tank games conducted by Dickinson and Shanks (Dickinson et al., 1984; Shanks 1985a; 1985b; 1986; 1987), the mines always existed as part of the context, and they were also

the main potential cause of the outcome (explosion). When a shell hit on the tank, the explosion of the tank might not only be attributed to the shell, but also either the mines, or both the mines and an effective shell. Additionally, the mines can also be considered as a necessary cause of the explosion (or "enabling condition" in Cheng and Novick's terminology (1992)), since the explosions (the outcome) always happened with them. Thus, for the subjects, the effectiveness of the mines, and also the shell, are rather ambiguous. In this case, the ΔP value of the shell should no longer be calculated by a simple ΔP rule, but becomes a rather complicated one, based on joint probability concepts (see Baker et al., 1989; Baker & Mercier, 1989).

A similar but improved video game was first employed by Baker et al.(1989), in which the outcome was no longer the explosion of the tank, but its safety when it traversed the minefield. One potential cause of the outcome in the game was camouflage. The camouflage was a colour change of the tank, and it was carried out by the subject pressing a computer button to shoot a special bullet on the tank and paint it. Thus, although the mines were still acting as contextual stimuli, they were not the cause for the outcome any more. They also could not be considered as the necessary cause or an enabling condition. The only cause is visible, and easy to identify. Using this improved game, Baker and his colleagues (1989) showed that the subjects' judgments were consistent with the ΔP rule at least in the one action and one outcome condition.

The main results of these experiments are not surprising ones, since they have been repeatedly observed before and after these experiments, by using different experimental procedures, by selecting different actions, and by showing different outcomes to different subjects (Alloy & Abramson, 1979; Allan & Jenkins, 1980; Baker et al., 1989; Shanks & Dickinson, 1987; Shanks 1985a; 1985b; 1986; 1987; Wasserman et. al., 1993). The most interesting problem

is how to explain these results. Do they mean that subjects really use the ΔP rule in their judgment processes?

Computational Models of Contingency Judgments

1). Probabilistic contrast model (Cheng & Novick, 1990a)

In the history of searching for the best explanation of covariations detection, several models or strategies have been proposed. For instance, as mentioned above, the cell A model, the cell A versus B model, and the cell A versus C model (see Kao & Wasserman, 1993), the ΔD and ΔP models (Inhelder & Piaget, 1958; Jenkins & Ward, 1965) have all been investigated. These models and others can be grouped into two categories at the computational level: statistical models and associative models. Two models which have attracted a lot of attention can be treated as the representatives of the two computational categories. As a statistical model, the probabilistic contrast was proposed recently by Cheng and Novick (1990a; 1992; also Melz, Cheng, Holyoak & Waldmann, 1993). This model was inspired in part by Kelley's (1967, 1973) ANOVA model of social attribution. Cheng and Novick noted that sometimes people showed judgment biases, because they failed to use some of the available information. In modifying Kelley's model, Cheng and Novick kept the three basic dimensions and the three kinds of information people may use in their judgments, but they emphasized the influence of the interaction among these dimensions on human' judgments.

This model has been used to explain main effect and interaction contrasts of multiple causes on the outcome of different contingencies. A main effect contrast, ΔPi, which specifies a cause involving a single factor i, is defined by the probabilistic contrast:

$$\Delta P_i = P_i - P_{\bar{i}} \quad (6)$$

Where P_i is the probability of the outcome occurring when factor i is present and $P_{\bar{i}}$ is the probability of the outcome occurring when factor i is absent. Evidently, for judgments in a single contingency task, this equation is identical to the ΔP rule proposed by Ward and Jenkins (1965).

In explaining the interaction of different factors in the judgment processes, a two-way interaction contrast, ΔPij, involving causal factors i and j, is defined as:

$$\Delta P_{ij} = (P_{ij} - P_{\bar{i}j}) - (P_{i\bar{j}} - P_{\bar{i}\bar{j}}) \quad (7)$$

P, as before, denotes the probability of cases in which the outcome occurs when the factors are present or absent.

With the possible exception of the outcome density bias, this model was successfully used to account for subjects' mean judgments in many cases because, as mentioned above, in single contingency tasks subjects' judgments consistently change with the ΔP values.

2). An associative model (Rescorla & Wagner, 1972)

As a representative from the associative category, the *Rescorla-Wagner model* (1972) was first proposed to account for variations in the effectiveness of reinforcement in animal conditioning. Dickinson et al. (1984) introduced it into the explanation of human judgment. They

(see Dickinson et. al., 1984; Shanks & Dickinson, 1987) suggested that human covariation judgments, like animal learning, is an associative learning process. They argued that the contemporary theories of animal conditioning might be extended to explain human judgments of covariations.

If we liken the outcome to a US, and the action to a CS, similar phenomena concerning CS and US covariations have been observed in Pavlovian conditioning experiments. For instance, when P(US|CS) is held constant, the degree of excitatory conditioning decreases systematically as P(US|~CS) increases (Rescorla, 1968). Similar experimental results have been found in negative and null contingency situations (Rescorla, 1968; Rescorla & Wagner, 1972). A model called Contingency Theory was advanced in the late 60's by Rescorla (1968). In this model, Rescorla suggested that the associative strength of single pair of CS and US was determined by a comparison between P(US|CS) and P(US|~CS). It can be considered as a parallel formula to the ΔP rule since it employs the same conditional probabilities to describe the relationship between two variables. This model was quite successful in predicting the consequences of exposing animals to various correlations between a CS and US, but it was not able to explain all the phenomena in conditioning. For instance, it could not address some interrelations between a US and two stimuli, causing the overshadowing phenomena in conditioning (see the following section for details). As a result, a new model was developed by Rescorla and Wagner (1972). The new model is powerful and able to explain many complex conditioning phenomena (e.g., Kamin, 1969; Mackintosh, 1971; Miles & Jenkins, 1973; Miller & Matzel, 1989; Wagner, 1969; 1971; Wagner, Logan, Haverlandt, & Price, 1968; also see Miller, Barnet & Grahame, 1995 for a review).

In this model, the associative strength between all stimuli, or all actions, and the outcome on a given trial *n* is the sum of the strength of each action (A, B ...):

$$Vsum_n = VA_n + VB_n + \ldots \quad (8)$$

The amount of associative change between an action A and the outcome on trial *n* is:

$$\Delta VA_n = \alpha\beta\,(\lambda - Vsum_{(n-1)}) \quad (9)$$

ΔVA_n represents the change in associative strength between the outcome and the action when the action is present on trial number n. $Vsum_{(n-1)}$ represents the total associative strength between the outcome and all actions accumulated in all trials up to the previous one. λ represents the maximum level of associative strength that the outcome can support. Following Rescorla and Wagner's recommendation, its value for tank games will be arbitrarily set to 1 for trials where the outcome is present and 0 when the outcome is absent. α corresponds to the salience of the action, while β is the salience of the outcome. Both parameters influence the rate of change in associative strength, and they are allowed to take on values between 0 and 1.

The associative strength of a given action with a given outcome cumulates over trials according to:

$$VA_n = VA_{(n-1)} + \Delta VA_n \quad (10)$$

According to these formulas, when an action and the outcome are repeatedly presented

together, $Vsum_{(n)}$ will follow a negatively accelerated growth and reach asymptote at the maximum level of associative strength that the outcome supports (λ). For example, on trial n, if an action is presented with the outcome and if λ is greater than $Vsum_{(n-1)}$, the difference between them is a positive value. Then ΔVA_n is also a positive value and the change of associative strength will be an increment. After this trial, the total association $Vsum_{(n)}$ between the action and the outcome is the sum of $Vsum_{(n-1)}$ and ΔVA_n. It is greater than $Vsum_{(n-1)}$. As the associative strength grows, the difference between λ and $Vsum_{(n)}$ decreases. If, on the next trial, the action and the outcome are presented together again, the increment will become less and if we continue this manner the increment will become less as trials proceed.

It is important to note that this model is different from statistical models. It cannot compute contingency at once based on the cumulative cell frequencies. However, it can compute associative values which covary with contingency and change progressively with each trial. When a group of trials are computed one by one, according to these formulas, even the order of presentation of the trials will influence the associative values.

Dickinson and Shanks (see Dickinson et al., 1984; Shanks, 1985a; 1885b; 1986; 1987; Shanks & Dickinson, 1987) viewed this associative model as the best account of information processes in short memory when people made contingency judgments. Computer simulations based on the model closely matched experimental results not only in single contingency tasks but also in double contingency tasks.

Double Contingency

Shanks (1986) employed a tank game similar to that used by Dickinson et al. (1984) and

added a new factor, a plane, which might drop bombs, as another potential cause of the tank's explosion. In addition to the mines, which can be considered as a contextual cause, each game in this experiment contained a double contingency where both the firing and the plane each had their own influence on the explosions. The subjects played four games. In two control games, the plane was not shown. In one of these control games, the contingency was positive, the ΔP value of firing was .50 (.5 - 0); in another, the contingency was null, the ΔP value of firing was 0 (.5 - .5). In the other two games, the ΔP value of firing was still equal to 0 (.5 - .5), but the plane was presented, and its presence was always accompanied with the tank's explosion. In one game, it appeared only on some trials when the subjects fired the shell; in another game it appeared only on the trials when the subjects did not fire. Since, in both games, the probability of an explosion outcome with the plane (P(O|plane)) equals 1 and both P(O|shell) and P(O|mines) are smaller than 1, then ΔP_{plane} must be greater than either ΔP_{shell} or ΔP_{mines}. Thus, the plane can be regarded as a more valid cause for explosion. Compared to the positive control contingency, the plane overshadowed the effectiveness of the shells when it appeared only on the trials where a shell was fired, that is, the effectiveness of the shells was underestimated. Also, subjects gave higher judgments to the shells when the plane appeared only without the firing. In this condition, subjects seemed to underestimate the effectiveness of the mines, and, consequently, judged the effectiveness of the shells as higher. The Rescorla-Wagner model was found to account for these results very well.

Using different tasks, other studies have reported similar overshadowing effects (e.g., Chapman & Robbins, 1990; Cheng & Mercier, 1990; Gluck & Bower, 1988; Shanks, 1984; 1990; 1991). For example, Shanks (1991) presented his subjects fictitious cases of patients, and asked

them to quantitatively rate how strong some of the symptoms were related to some of the diseases. Two kinds of case sets, contingent and noncontingent ones, were included in the experiment. For instance, in one of the contingent sets, the Compound Symptom AB signalled the presence of Disease 1, but Symptom B alone signalled the *absence* of the disease, whereas Symptom C alone signalled the presence of the disease. In the noncontingent set, Compound Symptom DE signalled the presence of Disease 2, Symptom E alone also signalled the presence of the disease, but Symptom F alone signalled the *absence* of the disease. The crucial comparison in this experiment is between the association rating given to Symptom A for Disease 1 and the association rating given to Symptom D for Disease 2. Shanks found that the mean association rating was higher for contingent cue (Symptom A for Disease 1) than for the noncontingent cue (Symptom D for Disease 2). The reason for this is that Symptom D was always presented with another good predictor (Symptom E) of Disease 2, whereas Symptom A was not. Thus, on a relative level, Symptom D was not valid as a predictor as Symptom A and its relation to the disease was estimated correspondingly lower. Once again, Shanks (1991) found that the Rescorla-Wagner model accounted for the results very well. Thus we can say that the Rescorla-Wagner set of equations accounted for subjects' judgments in double contingency very well.

However, the results obtained by Shanks (1991) can also be explained by a different model. Melz et al. (1993) used the probabilistic contrast model to do so. According to their explanation, in the processes of judgments people may select some of the data (*a focal set* in their terminology) from all the available information (called *universal set*) and then compute a probabilistic contrast based on the selected information. In Shank's disease experiment, if the universal set was used in computing the probabilistic contrasts, symptoms A and D would obtain

the same value. This is because P(disease 1|symptom A) would equal P(disease 2|symptom D) and P(disease 1|~symptom A) would equal P(disease 2|~symptom D), then the contrasts must be the same. However, Melz et al (1993) suggested that a focal set must be established to obtain a final contingency assessment, especially if some of the predictive stimuli cannot be evaluated in isolation. In Shanks' (1991) problem, the contingency value for symptoms C and F can be assessed immediately because they always occur in isolation. However, because B and E sometimes occur alone and sometimes in combination with A and D respectively, their contingency value must be conditionalized. Because A always occurs compounded with B, it must be conditionalized to B or assessed in the context of the focal set of all B trials. Similarly, D must be conditionalized to E. Thus, if only the information involving symptoms A and B are selected for disease 1, and only symptoms D and E are considered for disease 2, different contrasts will emerge. Based on these focal sets, the probability values of the two diseases given symptoms A or D respectively will be the same (P = 1.0), however, the probabilities of the two diseases without the symptoms will be different. Without symptom A, disease 1 has never been observed, so P(disease 1|~symptom A) equals 0; while disease 2 has been observed all the time without symptom D, so P(disease 2|~symptom D) equals to 1. Thus, based on the selected focal sets, the *relative* value of the contrast difference of A is 1 - 0 = 1, while the relative value of D is 1 - 1 = 0.

Baker et al. (1993) pointed out that none of the double contingency studies mentioned in the preceding paragraph had included all the possible four cells in the corresponding contingency tables. Thus the comparisons of these studies were not perfect. To include all cells in one experiment and compare more conditions directly, Baker et al. (1993) and Vallée-Tourangeau,

et al. (1994) used an improved tank game to conduct their experiments, in which two actions, a camouflage paint and a spotter plane, could potentially protect the tank in the minefield. Their experiments confirm the importance of the overshadowing effect, which was again observed in different conditions. For instance, it was found that when there were competing sources of protection for the tank, even if the strongest cause (plane) was not strong as that in Shanks' (1986) experiment, [e.g., ΔP_{plane} equals .8: (P(O|with plane) = .9, and P(O|without plane) = .1 in Baker et al.'s (1993) experiments], the moderate efficacy of the camouflage [$\Delta P_{camouflage}$ = .5: P(O|with camouflage) = .75, and P(O|without camouflage) = .25] was underestimated relative to its actual value. It was also found that a strong negative relationship between one cause and the outcome overshadowed the judgment on another moderate cause. Finally, two moderate correlations overshadowed each other.

Again, both the Rescorla-Wagner model and the probabilistic contrast model can account for the overshadowing phenomenon observed in double contingencies. Thus overshadowing studies cannot in, and of themselves, help to decide whether the associative mechanism or the probabilistic contrast is the better model to explain contingency judgments. In addition, the debate about which class of model provides the best account of contingency judgments has also indiscriminately pooled together the results of on-line and off-line studies. Unfortunately, as discussed before, the two paradigms do not appear to be equivalent. Thus, to differentiate the two classes of the models, it may be necessary to return to the basics and re-analyze in detail how single contingency judgments work.

On the surface, with the possible exception of the outcome density bias, both models seem to account well for single contingency judgments. For example, as mentioned before, several

studies have reported that human judgments of covariations are fairly close to the normative contingencies as calculated by the ΔP rule (Alloy & Abramson, 1979; Allan & Jenkins, 1980, 1983; Baker et. al., 1989; Dickinson et. al., 1984; Shanks, 1987; Wasserman et. al., 1993). These results can also be successfully explained by the Rescorla-Wagner model. Yet, as extensively as it has been studied, single contingency judgment has certainly not been investigated exhaustively. This argument is similar to the criticism made by Baker et al. (1993) about the double contingency conditions. For instance, when the ΔP value remains constant, different combinations of the four cells have not been all examined. In most previous studies, when the ΔP values were not equal to 1, extreme values like 1 or 0 have almost never been adopted for either P(O|A), or P(O|~A). To obtain a ΔP of .5, P(O|A) = .75 and P(O|~A) = .25 have often been used together; whereas P(O|A) = 1 with P(O|~A) = .5, or P(O|A) = .5 with P(O|~A) = 0 have rarely been used (but see Shanks, 1986; Wasserman et al., 1993). As a result, direct comparisons among these conditions have not been fully investigated. This is important because we do not know whether people's judgments will still be consistent with ΔP values in the situations where extreme values are used for either P(O|A), or P(O|~A) and any evidence of departure from ΔP in single contingency judgment would be difficult to handle by a probabilistic contrast model. In addition, the use of extreme values for probabilities, as well for frequencies is relevant to the logical and philosophical issues of sufficiency and necessity in causality.

Sufficiency and Necessity

As illustrated in Table 3, if P(O|A) = 1, cell A must be 1.0 and cell B in the corresponding contingency table (the frequency of non-outcome with the action) must equal zero.

Thus, the action becomes a sufficient cause to the outcome, since with it the outcome must occur.

Table 3

The contingency table when the action is a sufficient cause for the outcome.

	OUTCOME	NO OUTCOME
WITH ACTION	A	B = 0
WITHOUT ACTION	C	D

When P(O|~A) = 0, cell C (the frequency of outcome without the action) must equal zero. Then, the action becomes a necessary cause, since all the occurrences of the outcome must be accompanied by it. The corresponding contingency table is shown in Table 4.

Table 4

The contingency table when the action is a necessary cause for the outcome.

	OUTCOME	NO OUTCOME
WITH ACTION	A	B
WITHOUT ACTION	C = 0	D

In the domain of philosophy, a tradition dating back to Aristotle defines causality in terms of "necessity ceteris paribus": all other things remaining unaltered, if some condition p is

necessary for some outcome q, then q cannot occur unless p is satisfied. However, an alternative approach came from Mill (1843), who viewed causation in terms of "sufficiency ceteris paribus", a position supported by many others (e.g. Hempel, 1964; Popper, 1959): all other things remaining constant, if some condition p is sufficient for some outcome q, then q must occur if p is satisfied. If p is both necessary-and-sufficient for q, then q must occur if p is satisfied and cannot occur unless p is satisfied.

However, in the psychological literature, there have been few reports of judgment experiments related to the sufficient and/or necessary status of potential causes. The studies that were done all belong to the social and developmental areas (e.g., Bindra, Clarke, & Shultz, 1980; McGraw, 1987; Schustack & Sternberg, 1981; Siegler, 1976; Shultz, Butkowsky, Pearce & Shanfield, 1975; Shultz & Mendelson, 1975). Some of these studies compared people's judgments on multiple necessary causes (each of two or more factors is necessary for the outcome) and multiple sufficient causes (at least one of two or more factors is sufficient for the outcome). For example, Bindra et al. (1980) used four cards to show multiple necessary causes to the subjects. On the front of these cards four different combinations of two factors were shown: 1) a square shape with pink colour, 2) a square shape only, 3) pink colour only, and 4) no square, no pink colour. As an outcome, a drawing of a tree on the back of the card was predicted only by simultaneous presence of both a square shape and pink colour on the front. Thus, the pink colour and the square shape on the front of the card were the multiple necessary conditions for the appearance of a tree on the back. Another set of four cards showed multiple sufficient causes. A drawing of a house on the back was predicted by either a star shape *or* an orange colour on the front. Thus on three cards' backs there was a drawing of a house. Unfortunately, while this

study involved judgments for necessity and sufficiency, it did not include contingency judgments.

In addition, most of these social or developmental studies were conducted off-line, as well as connected with logical argumentation and verbal comprehension. For example, to examine the impact of sufficient and necessary causal conditions on the social attribution of blame for a negative outcome, McGraw (1987) used a series of statements and questions. One of these sets of statements is a basketball game scenario, in which losing is the negative outcome. For example: "Jim fouled out against the University of Michigan, the team lost" (evidence for sufficiency); "Jim fouled out against Michigan State, the team won" (evidence against sufficiency); "Jim did not foul out against the University of Iowa, the team lost" (evidence against necessity); and "Jim did not foul out against the University of Illinois, the team won" (evidence for necessity). After these statements were presented several times, the subjects were given other statements and asked to indicate the extent to which each of these statements was true on a nine point scale. The new statements used to check on necessity were: "Only when Jim fouls out does the team lose"; "If Jim does not foul out in the future, the team will not lose"; and so on. Some of the checks on sufficiency were: "Every time Jim fouls out, the team loses"; "If Jim fouls out in the future, the team will lose". In this study, the only information presented to confirm the sufficient condition was that of cell A, and for confirming the necessary condition was that of cell C. Also, the subjects' task was to evaluate the statement of either necessity or sufficiency. Since they relied on logic statements and proposition rules rather than presenting the contingencies directly, these off-line studies is also called deductive causal reasoning (see Cheng & Nisbett, 1993). They cannot differentiate the subjects' judgments on the necessary and/or the sufficient conditions from their ability to understand the logic statements and propositions per se.

An on-line experiment was conducted by Siegler (1976) to examine the developmental tendency of using sufficient and necessary information. Four conditions were shown to the children. In the sufficient and necessary condition, a light was turned on (outcome), after a card was inserted into a slot (action). This happened 6 times. In the necessary condition, the card was inserted 12 times, and the light came on 6 times among these trials. In the sufficient condition, the card was inserted 3 times and the light was turned on each time, but it also came on for another 3 trials. In the neither sufficient nor necessary condition, the card was inserted 8 times, and the light was turned on 6 times, but none of the lighting coincided with inserting the card into the slot. It was reported that, before the age of 8, children could not distinguish either the necessary or the sufficient cause from the cause which was both necessary and sufficient. Again, this on-line experiment did not show the information in all of the four cells of the corresponding contingency tables. For example, it did not show any information corresponding to cell D. For the both sufficient and necessary condition, only the information of cell A was presented, and for the necessary condition, only cells A and B were involved. On the other hand, for the neither sufficient nor necessary conditions, only cells B and C were presented.

This lack of use of some of the information cells was also found in other studies. For example, Bindra et al. (1980) argued that the cell A information was enough for the necessary and sufficient condition, cells A and B were enough for the necessary condition, cells A and C were enough for the sufficient condition, cells A, B and C were enough for neither necessary nor sufficient condition. These arguments leave cell D to be ignored, and these studies did not really deal with the contingencies at all. As Cheng and Nisbett (1993) criticized, contingency *is* important to decide the causal relationship between two events. In the deductive causal reasoning

tasks, which involved the necessity and sufficiency, the assumptions about contingency have been elicited. Why not do the same on-line? These studies ignoring some of the contingency information might be all based on the philosophic concepts of necessity and sufficiency, and contingency cannot be represented in propositional logic. In philosophy, if p is known to be a sufficient condition for q, then q can be predicted to occur whenever p occurs. However, the occurrence of q cannot be predicted when p does not occur. Thus, in the relevant contingency table, only cells A and B are presented. Actually, the sufficient cause may not be a cause for the outcome at all. Since if the frequency of cell D equals zero, then P(O|~A) will equal 1, and the contingency between this cause and the outcome will be zero. This means that no matter whether the cause is putatively presented or not, the outcome occurs, and there is no causal relationship between them. Similarly, in philosophy, if p is known to be necessary for q, then we can only predict the absence of q from p being absent. This has the same meaning as cell C equals 0 in the corresponding contingency table. But we cannot infer the real contingency between this necessary cause and the outcome. In their paper, Cheng and Nisbett (1993) suggested that the deductive causal reasoning tasks needed to involve the contingency. The other side of the coin is that the necessity and the sufficiency are also need to be presented in on-line contingency studies.

People's on-line covariation detection of putatively sufficient or necessary "cause" has not been directly investigated. Thus, the purpose of the thesis is twofold: (1) to compare on-line judgments relevant to the sufficient and/or necessary conditions, and (2) to evaluate different judgment rules, more specifically, to evaluate the relative merits of the probabilistic contrast model (Cheng & Novick, 1990a, 1992) and the Rescorla-Wagner model (1972) in single

contingency situations. This is achieved in four experiments systematically manipulating the frequencies of all cells in contingency tables on-line.

Experiment 1

In this experiment, the participants were asked to play eight video games similar to that used by Baker et al. (1989). The participants estimated, on-line, the extent to which a camouflage paint made a tank safe or unsafe as it passed through a minefield. The experiment was a 4 X 2 within-subject factorial design, involving four contingencies of camouflage (ΔP = .625, ΔP = .667, ΔP = .714, and ΔP = .769), and two "causal" statuses for these contingencies: sufficient and necessary. Each combination of ΔP value and "causal" status produced one of the eight games played by the subjects. In addition, the following controls were implemented. First, since previous studies suggested that the number of outcomes influenced subjects' judgments (See Dickinson et al., 1984), the frequency of the outcome was kept constant across all games. Second, for each game, the 40 trials were divided into two 20-trial periods, consisting of identical frequencies in each of the four cells of the contingency tables. To achieve this, the subjects had to be completely passive. During the game, the subjects did not paint the tank themselves. All the camouflage paints were produced by the computer program, and systematically displayed to the subjects. These controls guaranteed that within each game all subjects were shown the same frequencies in the four different cells, and the same contingencies.

All contingency tables for the 20-trial period of each game are shown in Table 5. The ΔP values of the camouflage paint for both Game 1 and Game 2 were equal to .625. In Game 1, this ΔP was obtained by subtracting 0 (P(safety | no camouflage)) from .625 (P(safety | camouflage)).

Table 5

Frequencies of Camouflaging Paint and Explosion in the 20 Trial Periods of Each Game in Experiment 1

	SAFE	EXPLODE
PAINT	10	6
NO PAINT	0	4

GAME 1 (necessary)
ΔP=.625-0 =.625

	SAFE	EXPLODE
PAINT	4	0
NO PAINT	6	10

GAME 2 (sufficient)
ΔP=1-.375=.625

	SAFE	EXPLODE
PAINT	10	5
NO PAINT	0	5

GAME 3 (necessary)
ΔP=.667-0=.667

	SAFE	EXPLODE
PAINT	5	0
NO PAINT	5	10

GAME 4 (sufficient)
ΔP=1-.333=.667

	SAFE	EXPLODE
PAINT	10	4
NO PAINT	0	6

GAME 5 (necessary)
ΔP=.714-0=.714

	SAFE	EXPLODE
PAINT	6	0
NO PAINT	4	10

GAME 6 (sufficient)
ΔP=1-.286=.714

	SAFE	EXPLODE
PAINT	10	3
NO PAINT	0	7

GAME 7 (necessary)
ΔP=.769-0=.769

	SAFE	EXPLODE
PAINT	7	0
NO PAINT	3	10

GAME 8 (sufficient)
ΔP=1-.231=.769

Thus, for the tank, the camouflage was a necessary cause of protection from explosion since the tank could not pass the minefield safely without it. In Game 2, the two corresponding conditional probabilities were 1 and .375. With these probabilities, the camouflage becomes a sufficient cause of protection since, with it, the tank would always pass the minefield safely. When the camouflage was a necessary cause, the tank would always explode without it; so the minefield was extremely dangerous. When the camouflage was sufficient, the minefield was proportionally less dangerous even if the simple frequency of the explosion without paint was higher (10 instead of 4).

The other ΔP values were, respectively, .667 in Games 3 and 4, .714 in Games 5 and 6, and .769 in Games 7 and 8. Odd game numbers correspond to the contingencies where the camouflage was a necessary cause of protection, whereas even game numbers correspond to sufficient ones.

Method

Subjects. Twenty-four undergraduate and graduate students from University of Ottawa participated in this experiment. They were each paid $5.00 for their participation.

Apparatus. An IBM compatible PC computer (Zenith) with a TTX CGA colour monitor was used to display the stimuli and record the data. The monitor's resolution was set to 320 X 200 pixels. The game program was written in Microsoft's Quick Basic.

Stimuli. The stimuli are depicted in Figure 1. The minefield was a 4.5 cm by 5.5 cm dotted cyan rectangle located in the upper left corner, 1.5 cm from the top of the screen. The camouflage paint was applied to the tank by a moving "paint bullet" launched from a U-shaped, cyan paint source. It was 0.8 cm by 1 cm and was located near the bottom of the screen,

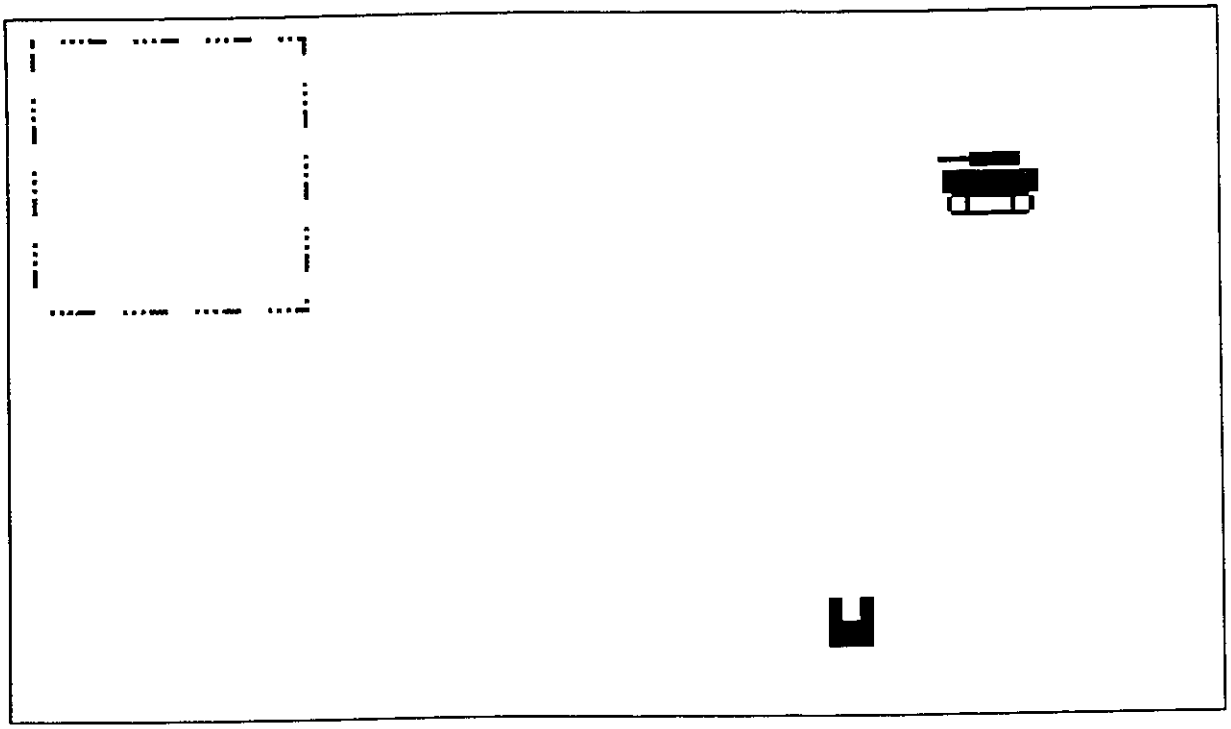

Figure 1. The tank, minefield, and paint source presented to the subjects in the experiments.

14.5 cm from the left-hand edge. The tank was pink with white tracks before it was painted. When it was painted, the body's colour changed from pink to cyan. The tank's dimensions were 1.6 cm by 1.3 cm. The colours of the tank, minefield, and paint source were kept the same through all eight games.

Procedure. The participants sat in front of the computer and read the 5 page instructions on the screen. After they finished reading one page of the instructions, they could press the space bar on the keyboard to read the next page. The exact instructions were as follows:

Page 1:

You will have a chance to play 8 games designed to let us know more about human judgement processes. In each game, 40 trials are included, and on each trial there is a tank and it is going to pass a minefield. The mines in the field are visually guided ones, which will explode and destroy the tank if they can see it. Around the centre bottom of the screen there is a Paint-O-Ray gun. In some trials for PROTECTING the tank it automatically shoots a special bullet and then changes the tank's colour from pink to BLUE.

Page 2:

If the COLOUR CHANGE makes the tank more difficult to see, it will be safer in the minefield. But the mines themselves may not be PERFECT. Even without the colour change, the tank may not explode. It is also possible that THE COLOUR CHANGE makes the tank easier to see. The tank could then explode more often. On each trial, before the tank enters the mine field, it stops and you will be asked to judge if the tank will explode or not in the minefield. After you make your judgment the tank will enter the minefield, and then you can find what happens to it. Whatever your judgment is, the tank's explosion is decided by the computer program.

Let's look at the tank, the minefield, and the gun.

(The participants saw the image depicted in Figure 1. After they pressed the space bar, Page 3 was presented.)

Page 3:

Before the game starts you will have 8 practice trials to familiarize yourself with

the game. In the practice trials, each condition will happen twice in a random order. The tank will explode four times: twice with colour change (tank's colour is BLUE), twice without colour change (tank's colour is PINK). The tank will also safely pass the minefield four times: with or without the colour change.

Page 4:

After the practice trials and after each 20 trials in the game you will be asked to assess the COLOUR CHANGE'S efficacy to protect the tank on a scale ranging from -100 to + 100.

+100 means that the colour change is perfectly SAFE. It camouflages the tank completely.

0 means that the colour change has NO EFFECT. The tank is neither more nor less visible when its colour is changed.

-100 means that the colour change is excessively DANGEROUS. It makes the tank perfectly visible and causes it to explode every time.

A score between 0 and 100 would mean that the colour change is partially able to make the tank safer.

You can now begin the practice trials.

(Then the subjects had 8 practice trials. The experimenter observed them during the practice, and answered their questions. After these practice trials the subjects were shown Page 5.)

Page 5:

Since you do not have any real information right now, you may begin the game by GUESSING the efficacy of the colour change to protect the tank.

> After the initial guess, your other judgments should be based on all the available trials in that game you have seen. As the trials go by, remember that the efficacy of the COLOUR CHANGE (from PINK to BLUE), and the MINE's ABILITY to see the tank remain consistent within every single game, but may CHANGE from one game to the next.

After the subjects read all of the instructions, they were asked to guess the efficacy of the painting. When the experimenter judged that the subjects understood the principle of the game, the subjects started the test games by themselves. The computer guided them through the games, but if they had any questions they could pause the experiment, and ask the experimenter in a neighbouring room. After each 20-trial period, the participants were asked to use the same scale to estimate the efficacy of the paint. Thus, they estimated the efficacy of the camouflage paint twice in each game. The whole experiment lasted about 60 minutes for each subject. All their estimates were automatically recorded by the computer program.

The order of presentation of the eight games was counterbalanced across the subjects. The order of appearance of the 20 trials in each period of any game was randomized.

Statistical analyses. In this and all subsequent experiments, .05 was selected as the criterion for tests of significance. Univariate F values are reported throughout. However, all within-subject effects were corroborated with multivariate tests. For each experiment, the various contingencies tested led to a profile of mean estimates to be compared with the normative profile. To take into account the three parts of the profile's similarity (Blashfield and Aldenderfer, 1988; Cronbach and Gleser, 1953), three particular operations were performed. First, to insure uniformity of size or level, we scaled all judgments and theoretical values to the same range

(-100 to +100). Second, to assess the similarity between the shape of the model and the shape of the data (pattern of dips and rises across means), we correlated (*Pearson's r*) the theoretical means with the empirical means of estimates. In addition, in order to find out whether any alternative model provided significant shape information above and beyond the normative value, multiple regression was performed and ΔP was forced to enter the equation first, then the other models. This is a stringent test because it goes beyond testing for mere equivalence of a given model and the normative value. Third, to assess the distance between the theoretical means and the data, the standard deviation of the difference between theoretical and data points was calculated (square root of: the sum of squared differences between theoretical means and empirical means divided by the number of means) and compared across models with a repeated measure ANOVA. The best model should simultaneously have the least error (smallest standard deviation of difference) and the best pattern similarity (largest correlation). These tests were conducted on the means instead of on the individual data points because the theoretical models make predictions globally and do not include mechanisms for individual modulation.

Results and Discussion

The average initial guess for the efficacy of the camouflage was 46.15. Since statistical analyses showed no significant difference between middle judgments (after 20 trials) and final judgments (after 40 trials), only final judgments are represented graphically. The participants' final estimates of the paint efficacy for each game in the present experiment are shown in Panel A of Figure 2. This figure shows clear judgment patterns. First, the figure showed that the actual contingencies influenced the estimates, since the estimates consistently increased in both the

sufficient and the necessary condition as the ΔP value increased. This portion of the results replicated the prior findings of Baker et al. (1989), Dickinson et al. (1984), and Shanks (1985b). Second, the average estimates varied according to the necessity or the sufficiency of the camouflage. The sufficient camouflage was always estimated more effective than the corresponding necessary one.

The statistical analyses support these impressions. A three-factor, within-subject, analysis of variance was carried out on the four contingencies (.625, .667, .714, .769), two conditions (necessity/sufficiency), and two judgments (20/40 trials). Significant main effects were found for the contingency, $F(3, 69) = 13.58$, $MSE = 496$, $p < .01$; and for the condition, $F(1, 23) = 38.71$, $MSE = 1082$, $p < .01$. No significant main effect was found between the judgments after 20 and after 40 trials. No interaction was significant.

The results of the present experiment, again, indicated that human judgments of covariations in a single contingency task are related to the contingencies between the outcome and the action. However, the crucial finding of the present experiment is that people judged the sufficient and the necessary factors differently. To test the explanatory power of the various contingency rules, they were evaluated individually. The cell A rule definitely could not be used to account for the subjects' judgments, since, when the frequency of cell A was higher, the mean judgment did not become larger. For instance, in Game 1, cell A had higher frequency (10) than in Game 2 (4) but subjects gave higher judgment in Game 2. Also, Games 1, 3 , 5, and 7 had the same frequencies (10) in cell A, but the mean judgments were different on these games. The consistent application of either the "cell A versus cell B" rule or the "cell A versus cell C" rule does not explain subjects' mean judgments. For the same contingency games, the frequency

differences between cells A and B, or cells A and C in the necessary conditions (Game 1, 3, 5, or 7) were larger or at least identical to that of the corresponding sufficient games (Game 2, 4, 6, or 8), but the mean judgments were always lower in the necessary conditions than in the corresponding sufficient ones. A similar conclusion applies for the ΔD rule, since according to it judgments should remain constant for a fixed contingency and yet the mean judgments were different between the sufficient and the necessary conditions. Although the ΔP rule can account for the judgment differences among different contingencies, it cannot successfully explain the judgment differences observed between the necessary and the sufficient conditions. When extreme values were used for one of the two conditional probabilities, people considered the sufficient cause as more effective than the necessary one. Thus, people's judgments did not rely directly on the ΔP value. In other words, in single contingency tasks, people do not appear to make their judgment based on the probabilistic contrast model.

In Figure 2, we can observe that the average estimates in the sufficient condition are higher than the actual contingencies of the corresponding games, whereas, in the necessary condition, the estimates are lower than the contingencies. To test if the subjects' estimates in necessary and/or sufficient condition represent significant judgment biases, the difference between the real ΔP values (Panel B in Figure 2) and the subjects' estimates in either condition were tested against zero (*paired t-tests* with a Bonferroni correction for the number of comparisons). Compared with the real ΔP values, the subjects reliably underestimated the efficacy of the camouflage in the necessary condition, $t(95) = -4.28$ ($p < .01$), whereas they overestimated the efficacy in the sufficient condition, $t(95) = 6.38$ ($p < .01$). Thus, compared to the normative values, the judgments were significantly biased in both the sufficient and the necessary

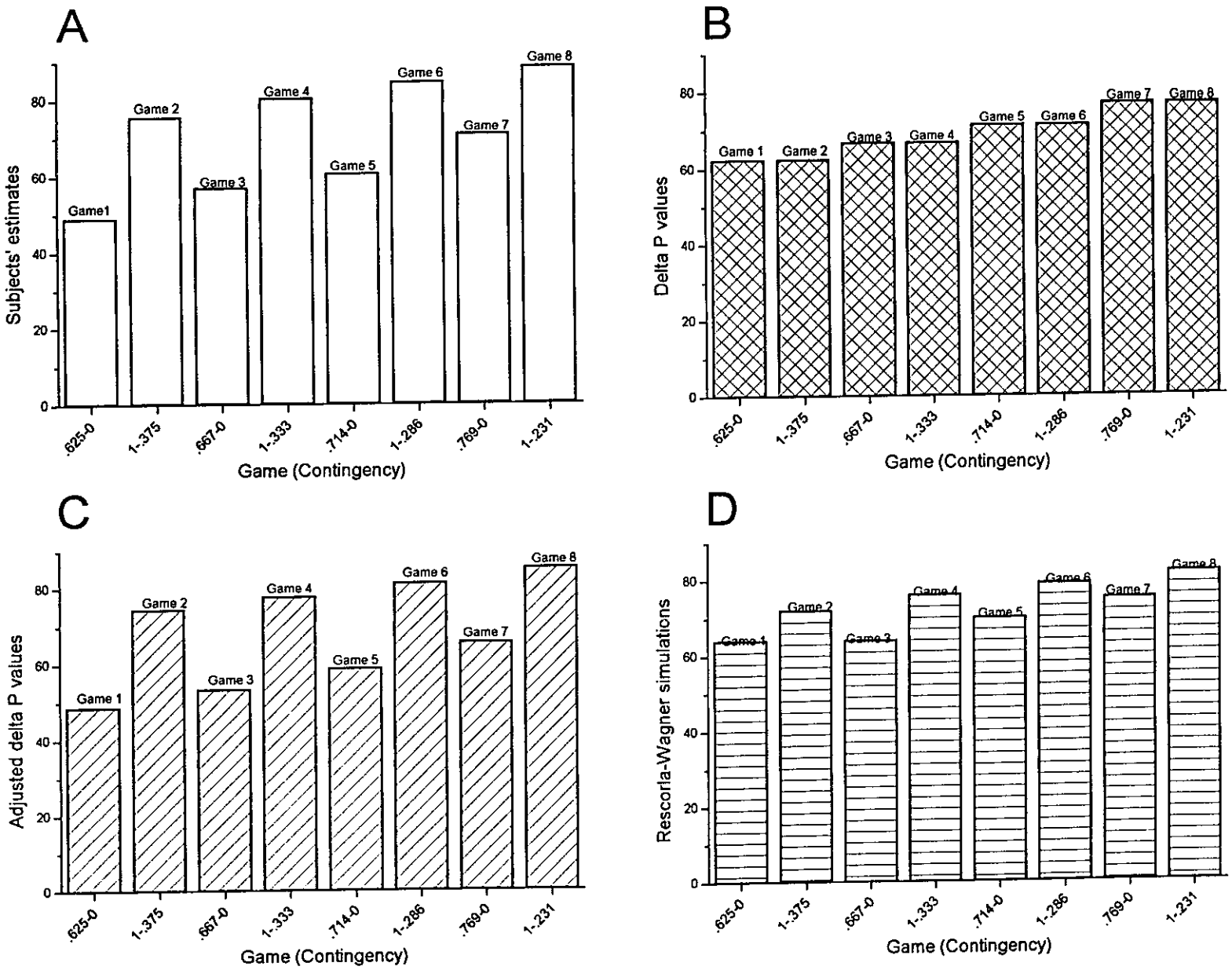

Figure 2. Subjects' mean judgments in Experiment 1 and the calculated values from the delta P rule, the adjusted delta P rule, and the Rescorla-Wagner model.

conditions. These judgment biases have not been reported before in on-line experiments.

The predictions of the Rescorla-Wagner model were simulated for all the games in the present experiment. During each game, the subjects were actually shown 40 trials. Thus, in each run of the simulation, exactly 40 trials were processed in random order according to the formula. There were the same numbers of different types of simulated trials as the participants experienced during the different games. For each game, 100 runs were performed and averaged. The α parameter was set to 1 for the paint, and to .1 for the context (without the paint). β was .35 for the safe trials (outcome), and .3 for the unsafe ones (no-outcome). These parameters were selected by a systematic (increments of .05 for each parameter individually) but non-exhaustive search of the parameter space. This search was first performed using a custom program and later confirmed using a published tool (Mercier, in press). With these parameters, the model produced the simulation pattern as shown in Figure 2 (Panel D). A visual inspection of the results of the simulation shows that, in general, the simulated values for these games are much closer to subjects' real judgments than those obtained from the ΔP rule. The larger the judgment value becomes, the larger the simulated value is. Also, the simulated values correctly show the difference between the necessary and the sufficient conditions: they were always larger in the sufficient condition. Still this model did not simulate all the subjects' estimates perfectly. For example, it predicted that the subjects should make a higher judgment in Game 7 (75) than that in Game 2 (72), but the reverse pattern was observed (76 for Game 2 and 71 for Game 7). Also, the participants judged Game 1 and Game 3 differently (49 for Game 1 and 57 for Game 3), yet the simulated values showed no difference between them (64 for both games). Finally, for some games (e.g., Game 1, Game 3, and Game 5) the subjects did not make estimates as high as those

produced by the simulation.

Although imperfect, the simulation results suggest that an associative model, such as the Rescorla-Wagner model (1972) fits the pattern of single contingency judgments much better than a statistical model, such as Cheng and Novick's (1992) normative probabilistic contrasts. However, if a non-normative or weighted version of the ΔP rule is considered, then a new picture emerges. For instance, Wasserman, Dorner, and Kao (1990) studied which cell in the contingency table is perceived as the most important and the least important by people making their judgments. Wasserman et al. suggested that the order of the importance is: cell A > cell B > cell C > cell D. Not limiting ourselves to this strict pattern, we tried various weights on each cell until we obtained a more satisfactory fit between the subjects' mean estimates and empirically adjusted ΔP values. The best and the simplest adjustment found was:

$$Adjusted\ \Delta P = A/(A+1.75B) - C/(C+1.75D) \qquad (11)$$

The difference between this formula and the ΔP rule is that a weight is added to both cell B and cell D. Figure 2 also shows the values calculated from this formula (Panel C). From this figure, we can compare the participants' estimates to the calculated values from the ΔP rule, the adjusted ΔP rule, and the simulated values from the Rescorla-Wagner model. Visual inspection clearly shows that the calculated values from the adjusted ΔP rule coincide well with the actual estimates. The only difference is that the calculated values are slightly lower than the actual final estimates in most of the games.

The largest standard deviation of the differences (SD_{diff}) was obtained between the

subjects' mean estimates and the ΔP (16). The smallest SD_{diff} was produced by the adjusted ΔP rule (3). The Rescorla-Wagner model produced a value falling between those two rules (8). Thus, for this experiment, the associative model is more accurate than the probabilistic contrast (ΔP). A repeated measure analysis of variance confirmed that there were some significant differences among the models, $F(2, 14) = 12.65$, $MSE = 2582$, $p < .01$. The mean SD_{diff} for the Rescorla-Wagner model was significantly smaller than that of ΔP (Tukey) and so was the mean error for the adjusted ΔP. The difference between the Rescorla-Wagner model and the adjusted ΔP was not significant.

In terms of pattern of mean predictions, the adjusted ΔP and the Rescorla-Wagner models ranked similarly. The adjusted ΔP rule had the largest correlation coefficient with the subjects' judgments ($r(6) = .99$; $p < .01$) with the Rescorla-Wagner model closely behind ($r(6) = .95$; $p < .01$). The correlation between the ΔP rule and the estimates was the smallest and not significant, $r(6) = .50$. In addition, multiple regressions forcing the normative (ΔP) predictions in the equation first, and adding an alternative model second showed that both the adjusted ΔP rule and the Rescorla-Wagner model added significant information above and beyond the norm. The adjusted ΔP rule could account for an added 76 percent of the variance in the subjects' estimates (R^2 for ΔP and adjusted ΔP both in the regression equalled .99), whereas, the Rescorla-Wagner model accounted for an added 72 percent of the variance (total $R^2 = .95$). ΔP alone accounted for only 23 percent of the variance and did not contribute significantly to the regression equation. These results, again, suggested that the ΔP rule is not a good model to explain the data obtained in the present experiment. However, both the adjusted ΔP rule and the Rescorla-Wagner model showed strong explanatory power with the adjusted ΔP appearing somewhat more accurate.

From Experiment 1 alone, it is not possible to determine if the over-estimation of the sufficient factor and the under-estimation of the necessary factor is constant. It is also not possible to determine if the over/under-estimation split is strictly related to the necessity and sufficiency of the cause, because Experiment 1 only contained conditions involving extreme values. Experiment 2 attempts to clarify this by comparing conditions without extreme values.

Experiment 2

Experiment 2 was designed to test if subjects always overestimate the efficacy of the camouflage in the sufficient condition, and if they always underestimate the efficacy in the necessary condition, by studying contingency tables without zero frequencies. This experiment was also designed to test the explanatory power of the ΔP rule, the Rescorla-Wagner model, and the adjusted ΔP rule for these conditions. The experiment used the same kind of computer program and involved some of the games with extreme values used in Experiment 1. In addition, control conditions involving no extreme probabilities but having the same ΔP values were included.

Method

Subject and apparatus. Twenty-four subjects were recruited from University of Ottawa. They were either graduate or undergraduate students. They were each paid $5.00 for their participation. The apparatus and the stimuli described in Experiment 1 were used.

Procedure. The design was a within-subjects factorial arrangement involving 2 contingencies, 4 conditions. Thus, each subject was presented with eight 40-trial games and they were asked to make twice judgments for each game. Two camouflage contingencies, ΔP = .625 and ΔP = .71 (rounded number), were displayed in four different conditions: the sufficient condition (Games 4 and 8), the necessary condition (Games 1 and 5), and two control conditions for each of the necessary and the sufficient games (Games 2, 3, 6 and 7). In the two control conditions, no extreme value was used for either of the two conditional probabilities. The two probabilities for the different control conditions are different. For the eight games, the frequencies in each cell of the corresponding contingency tables were the same for all subjects, and they were

Table 6

Frequencies of Camouflaging Paint and Explosion in the 20 Trial Periods of Each Game in Experiment 2

	SAFE	EXPLODE
PAINT	10	6
NO PAINT	0	4

GAME 1 (necessary)
ΔP=.625-0=.625

	SAFE	EXPLODE
PAINT	9	3
NO PAINT	1	7

GAME 2 (control)
ΔP=.75-.125=.625

	SAFE	EXPLODE
PAINT	7	1
NO PAINT	3	9

GAME 3 (control)
ΔP=.875-.25=.625

	SAFE	EXPLODE
PAINT	4	0
NO PAINT	6	10

GAME 4 (sufficient)
ΔP=1-.375=.625

	SAFE	EXPLODE
PAINT	10	4
NO PAINT	0	6

GAME 5 (necessary)
ΔP=.714-0=.714

	SAFE	EXPLODE
PAINT	9	2
NO PAINT	1	8

GAME 6 (control)
ΔP=.818-.111=.707

	SAFE	EXPLODE
PAINT	8	1
NO PAINT	2	9

GAME 7 (control)
ΔP=.899-.182=.707

	SAFE	EXPLODE
PAINT	6	0
NO PAINT	4	10

GAME 8 (sufficient)
ΔP=1-.286=.714

again divided equally into two 20-trial periods. For all games, the number of outcomes was kept constant (10). The contingency tables for the 20-trial period of each game are shown in Table 6. The instructions were the same as described for Experiment 1. The participants were asked to judge the efficacy of the camouflage after each 20-trial of the games on a scale ranging from -100 to +100. The order of the games was counterbalanced across subjects, and the order of presenting the trials in each 20-trial-period was randomized.

Results and Discussion

The mean guess for the paint equalled 39.17. In Figure 3, Panel A depicts the means of the final judgments of camouflage efficacy for each of the eight games. The judgment patterns are similar to those in Experiment 1. The higher the contingency was, the higher was the estimate. Again, when the camouflage was a sufficient cause, the subjects judged it more effective than when it was a necessary one. For the two control conditions, the subjects' judgments fell between the estimates in the sufficient and in the necessary conditions.

The statistical analyses confirm these assertions. A 2 X 4 X 2 within-subject analysis of variance for the contingencies, the conditions, and the judgment (after 20 and after 40 trials) showed that the subjects could discriminate different contingencies, $F(1, 23) = 5.29$, $MSE = 538$, $p < .05$; and the conditions, $F(3, 69) = 19.51$, $MSE = 842$ $p < .01$. There was no significant difference between the judgments after 20 trials and after 40 trials. No significant interaction was found. Subsequent analyses were performed among the four conditions. Tukey's HSD test showed that the judgments in the sufficient condition (Games 4 and 8), which contained an extreme value on the probability of the outcome with the camouflage, were significantly different from those in the necessary condition (Games 1 and 5), which contained an extreme value on the probability

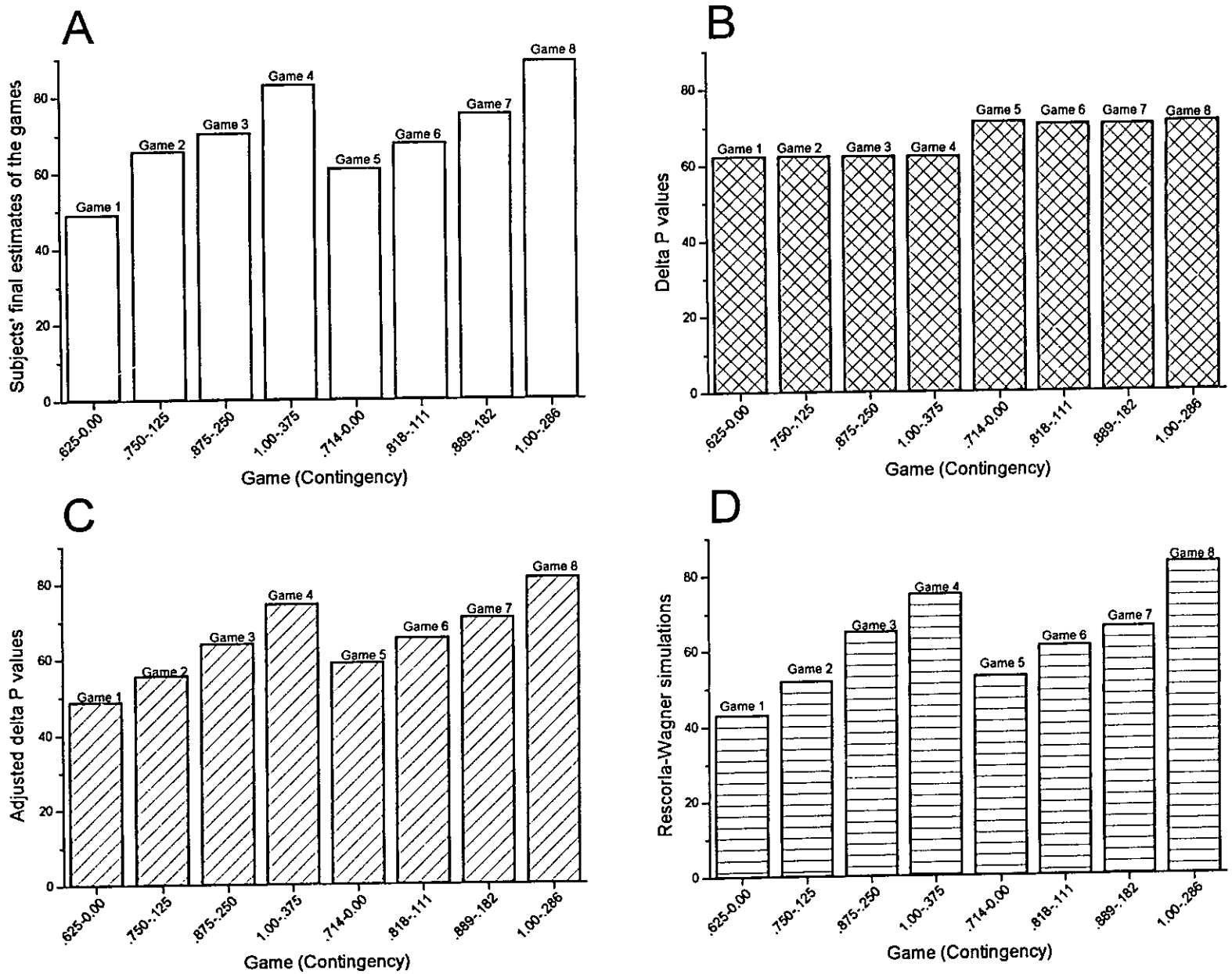

Figure 3. Subjects' mean estimates in Experiment 2 and the calculated values from the delta P rule, the adjusted delta P rule, and the Rescorla-Wagner model.

of the outcome without the camouflage; and also from either of the two control conditions (Games 2, 3, 6, and 7), which did not involve any extreme values. The difference between the necessary condition and either of the two control conditions was also significant. However, the judgment difference between the two control conditions was not significant.

Paired t-tests, with a Bonferroni correction, showed significant differences between the estimates and the actual contingency values in both the sufficient condition ($t(47) = 7.95, p < .01$) and the necessary condition ($t(47) = -3.59, p < .01$). However, the estimates for both of the control conditions were not significantly different from the actual ΔP values.

These results showed that people made graded judgments as a function of sufficiency and necessity. They overestimated the efficacy of the camouflage when it was a sufficient cause, and underestimated it when it was a necessary one. Contingency judgments based on tables containing no extreme value represented a gradation between those two extremes.

Figure 3 presents the calculated values from the ΔP rule (Panel B) and the adjusted ΔP rule (Panel C). The first impression given by a visual inspection of Figure 3 is that the ΔP values are those which match the estimates the least adequately. There are only two calculated ΔP values whereas the participants produced four different ones for each contingency. To assess which theoretical model can best account for the overall patterns of judgments, the same methods reported in Experiment 1 were used again. The standard deviation of the difference between the ΔP values and the mean estimates across the eight games was 11, and the correlation between the two series of data was low ($r(6) = .27$) and not significant. Using the same weights as in Experiment 1, the adjusted ΔP rule produced values which coincided well with the actual judgments. However, one weakness of these calculated values is that they are all lower than the

real judgments. The standard deviation of the differences between the adjusted ΔP and the judgments was 6, about half the size of the deviation from the ΔP rule. The adjusted ΔP correlated highly and significantly with the estimates ($r(6) = .97$, $p < .01$). Thus, in spite of the constant difference making all adjusted ΔP lower than the estimates, the weighted values fit the pattern of judgment means well.

When α and β values were selected to be the same as in Experiment 1, by using the standard simulator (Mercier, in press), the Rescorla-Wagner model was not as successful as reported by others (Dickinson et al., 1984; Baker et al., 1993; Shanks, 1987; 1991; 1993; Vallée-Tourangeau et al., 1994; Wasserman et al., 1993). These results are *not* shown in Figure 3 (but see below). The simulation could account for the difference between contingencies. It also explained the judgment difference between the sufficient and the necessary condition reasonably. However, it showed a tendency to understate that difference. For the two games in the sufficient condition, the simulated values were lower than the real judgments. On the other hand, for the games in the necessary condition, the simulated values were higher than the judgments. This means that the simulations decreased the judgment difference between the necessary and the sufficient conditions. To some extent, it produced a judgment pattern similar what the ΔP rule did.

It was also found that the simulation did not show enough difference between the sufficient condition and the control condition. For example, the simulated values did not show any difference between Game 3 and Game 4, and showed little difference between Game 7 and Game 8. Empirically however, the subjects judged the efficacy of the paint in the sufficient condition as significantly different from that in the control conditions. Similarly, the simulated

values for Game 1 and Game 2 showed little difference while the mean judgments were significantly different between the two games. Overall, the Rescorla-Wagner model with parameters (α_{paint} = 1, $\alpha_{context}$ = .1, $\beta_{outcome}$ = .35, and $\beta_{no\text{-}outcome}$ = .3) had a standard deviation of the difference from the estimates of 8 points. It correlated with the estimates ($r(6) = .83$, $p < .05$) at a level intermediate between the ΔP rule and the adjusted ΔP rule.

In simulating the associative model, the α parameter can be set at different values for the trials with the factor, and for the context alone. Similarly, the β parameter can be set differently for the trial with the outcome and without the outcome. Usually, the β parameter for a trial with the outcome is set to a larger value than that without the outcome on the grounds that the presence of an outcome is more salient than its absence (see Rescorla & Wagner, 1972, p86; also see Shanks, 1991, p434). However, this needs not always be the case. In the current set of experiments for instance, it could be argued that a safe outcome (absence of explosion) is *less* salient than an unsafe outcome (presence of explosion). Thus, several combinations of parameters other than those used in Experiment 1 were tested. It was found that if β for the outcome was equal to or larger than that without the outcome, the simulated judgment pattern could not be similar to the subjects' judgments. The simulations were always lower than the judgments in the sufficient conditions, and always higher than the judgments in the necessary conditions. Within a range, when β without the outcome was larger than that with the outcome, the model could produce a judgment pattern very similar to that of the participants' mean estimates. By using a custom program, a systematic search found the best parameters: α_{paint} = 1, $\alpha_{context}$ = .15, $\beta_{outcome}$ = .3, and $\beta_{no\text{-}outcome}$ = .65. The simulated values based on these parameters are shown in Panel D of Figure 3. This simulation produced final judgments similar to the adjusted ΔP rule values.

The standard deviation of the difference between these simulated values and subjects' mean judgments was 8. This value is the same as the first simulation reported above. It does not differ significantly from either ΔP or adjusted ΔP just as these two do not differ from one another ($F(2, 14) = 1.68, MSE = 9862$, n.s.). The correlation fits better ($r(6) = .98, p < .01$) than the first model did. As can be seen in Figure 3, the values obtained from the latter simulation were consistently smaller than the subjects' judgments, yet the pattern of differences among means is well reproduced.

When these new parameters were used to re-simulate the final judgments in Experiment 1, the simulated pattern was also very good. SD_{diff} grew larger (10), but the correlation improved ($r(6) = .97, p < .01$). The simulated values did show a tendency to be larger than the estimates for the sufficient conditions, and lower than the estimates for the necessary conditions. This might have caused the larger standard deviation of the differences.

Multiple regressions indicated that when ΔP entered the formula first, it could only account for 7 percent of the variance (not significant) of the mean estimates. The adjusted ΔP significantly accounted for an additional 89 percent more of the variance. The simulated values from the Rescorla-Wagner model with the best set of parameters also significantly accounted for an added 89 percent of the variance.

Overall, both the Rescorla-Wagner model and the adjusted ΔP rule can explain the subjects' judgments of Experiments 1 and 2 very well. It is also more and more evident that the ΔP rule is not a good model to explain the mean judgments of a group of people in single contingency tasks. Even as a descriptive model, it is not accurate at reproducing the biases observed in the subjects' mean judgments.

Experiment 3

In Experiments 1 and 2, judgment biases were found in both the necessary and the sufficient conditions. As mentioned before, when the camouflage is a sufficient cause, the probability of the outcome with the camouflage equals 1; when the camouflage is a necessary cause, the probability of the outcome without the camouflage equals 0. In the corresponding contingency tables, either cell B or cell C equals 0. There are two other possible conditions involving zero frequencies in the contingency table have not been dealt with in the first two experiments: either cell A or cell D could equal 0.

If cell A equals 0, the probability of the outcome with the cause (P(O|A)) is 0. If at the same time, the probability of the outcome without the cause (P(O|~A) is not 0, we have a negative contingency. When cell D equals 0, P(O|~A) equals 1. If the P(O|A) is smaller than 1, another negative contingency will be obtained.

Shultz, Butkowsky, Pearce & Shanfield (1975) reported some experimental results on children's causal judgments that are relevant to the sufficient and necessary conditions reported above. In their experiment, all the possible extreme values involving 0 in one of the four cells, were included. If cell A, and only cell A, equals 0, the cause and the outcome never happen together. They named this kind of cause an "inhibitory sufficient cause". They called the cause in the usual sufficient conditions, as studied in Experiments 1 and 2, a "facilitory sufficient cause". When cell D, and only cell D, equals 0, the outcome is always present without the putative cause. Although when the cause is given, the outcome is sometimes present, all the absences of the outcome must be accompanied with the cause. In this condition then, the cause can be treated as a necessary one for preventing the outcome, and it was called an "inhibitory

necessary cause" by Shultz et al. (1975), whereas the cause in the usual necessary condition was called a "facilitory necessary cause".

Experiment 3 was conducted to obtain judgment patterns using extreme values in negative contingencies. It was used to examine whether people judge the inhibitory necessary and sufficient conditions the same or differently than they did the facilitory conditions of Experiments 1 and 2. In previous research, it was reported that the ΔP rule could explain the judgments of the negative contingencies quite well in single contingency task (e.g., Baker et al., 1989; Dickinson et al., 1984).

The design was a 2 X 2 X 2 factorial arrangement involving two directions of the contingencies (positive and negative), two levels of contingencies in each direction (+/-.625, +/-.714), and the two conditions (sufficient and necessary). The inclusion of the positive contingencies was used to replicate the results obtained in the first two experiments while allowing for within-subject comparisons between the facilitory and the inhibitory conditions.

Method

Subjects and apparatus. Twenty-four students were recruited from the University of Ottawa. They were paid $5.00 each for their participation. The apparatus and the tank game were the same as described in Experiment 1.

Procedures. The same instructions were used as in Experiment 1. Again, each subject played eight games. Each game included 40 trials divided into two 20-trial periods. By controlling the frequencies in each cell of the corresponding contingency tables, all subjects were presented with identical contingencies. The number of the outcomes in all games was kept constant (10).

Table 7

Frequencies of Camouflaging Paint and Explosion in the 20 Trial Periods of Each Game in Experiment 3

	SAFE	EXPLODE
PAINT	6	10
NO PAINT	4	0

GAME 1 (inhibitory necessary)
ΔP=.375-1=-.625

	SAFE	EXPLODE
PAINT	0	4
NO PAINT	10	6

GAME 2 (inhibitory sufficient)
ΔP=0-.625=-.625

	SAFE	EXPLODE
PAINT	4	10
NO PAINT	6	0

GAME 3 (inhibitory necessary)
ΔP=.286-1=-.714

	SAFE	EXPLODE
PAINT	0	6
NO PAINT	10	4

GAME 4 (inhibitory sufficient)
ΔP=0-.714=-.714

	SAFE	EXPLODE
PAINT	10	6
NO PAINT	0	4

GAME 5 (facilitory necessary)
ΔP=.625-0=.625

	SAFE	EXPLODE
PAINT	4	0
NO PAINT	6	10

GAME 6 (facilitory sufficient)
ΔP=1-.375=.625

	SAFE	EXPLODE
PAINT	10	4
NO PAINT	0	6

GAME 7 (facilitory necessary)
ΔP=.714-0=.714

	SAFE	EXPLODE
PAINT	6	0
NO PAINT	4	10

GAME 8 (facilitory sufficient)
ΔP=1-.286=.714

The contingency tables for the 20-trial period of each game are shown in Table 7. Four contingencies, two positive (ΔP = .625, and ΔP = .714) and two negative (ΔP = -.625 and ΔP = -.714) were presented in either the sufficient or the necessary condition. Thus, the positive contingencies were presented either as a facilitory sufficient condition or as a facilitory necessary condition, while the negative contingencies were presented either as an inhibitory sufficient condition or as an inhibitory necessary condition. The subjects were asked as in the previous two experiments to judge the efficacy of the camouflage to protect the tank after each 20 trials on a -100 to +100 scale. The order of presentation of the games was counterbalanced across subjects. The order of display the trials was randomized within each 20 trial period.

Results and Discussion

The mean guess of the efficacy of camouflage was 29.17. The subjects' average estimates after 40 trials of each game are shown in Panel A of Figure 4. The estimates on positive contingency games paralleled the judgment pattern obtained in the first two experiments. As the contingency value became higher, the participants increased their judgment, and the sufficient camouflage paint was judged more effective than the necessary one.

For the negative contingencies, the participants reported negative judgments. Two aspects of the negative judgments mirrored the pattern of the positive ones. First, the subjects' judgments were consistent with the contingencies. When the contingency was more negative, the judgment was lower. Second, the subjects seemed to judge the inhibitory sufficient condition and the inhibitory necessary condition differently. However, their estimates were more negative in the sufficient conditions than in the corresponding necessary ones. The participants seemed to think the camouflage paint was more dangerous in the inhibitory sufficient condition.

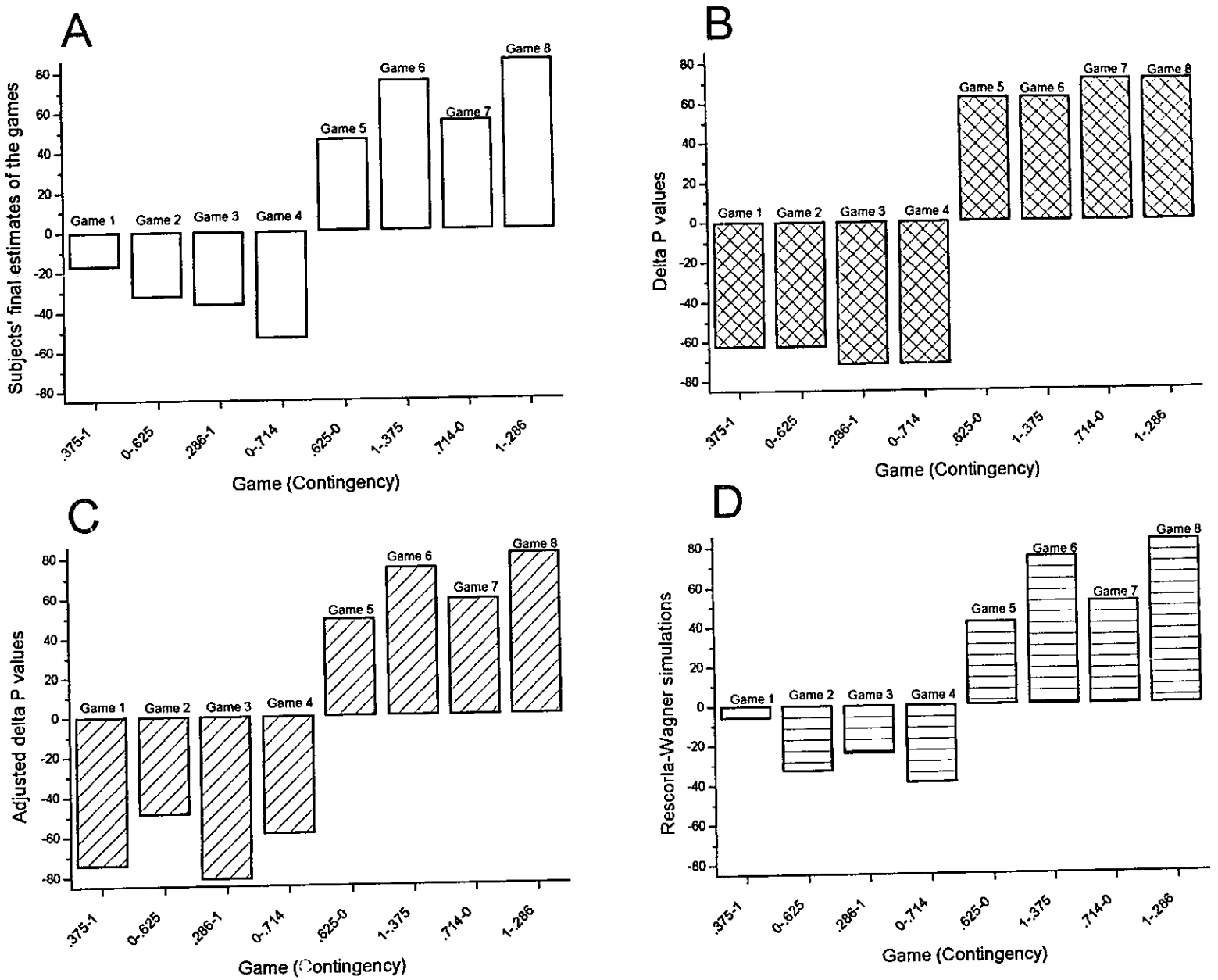

Figure 4. Subjects' mean estimates in Experiment 3 and the calcualted values from the delta P rule, the adjusted delta P rule, and the Rescorla-Wagner model.

The statistical analyses provided partial support for these assertions. An initial 2 contingencies (.625 or .714) X 2 signs (positive or negative) X 2 conditions (necessary or sufficient) X 2 judgments (20 or 40 trials) analysis of variance revealed significant differences between signs ($F(1, 23) = 169.74$, $MSE = 5897$, $p < .01$) and between conditions ($F(1, 23) = 5.64$, $MSE = 1857$, $p < .05$). The differences between the contingencies and between the trials were not significant. Significant interactions between the signs and the conditions ($F(1, 23) = 16.95$, $MSE = 2075$, $p < .01$) and between the signs and the contingencies ($F(1, 23) = 9.92$, $MSE = 1277$, $p < .01$) were also found. This direct analysis of the data makes it difficult to interpret the significant interactions and the relevant main effects. Even if the positive and the negative contingencies were perfect mirror images of one another, any other effect in the model would be likely to interact with the signs of the contingencies. To simplify the interpretation, the sign of the judgments in the negative games were all inverted. Thus, we treated the data as if all the contingencies were positive but presented in different contexts. The transformed data were re-analyzed according to the same plan as before.

Significant main effects were found for contingency ($F(1, 23) = 9.92$, $MSE = 1277$, $p < .01$) and condition ($F(1, 23) = 16.95$, $MSE = 2075$, $p < .01$). A reliable difference was also found between the positive contingencies and the transformed data from negative contingencies, $F(1, 23) = 19.38$, $MSE = 3488$, $p < .01$. This suggests that the judgments on the negative contingencies are not exactly a mirror pattern of the positive ones. Additionally, the sign of the contingencies interacted with the sufficient/necessary condition, $F(1, 23) = 5.64$, $MSE = 1857$, $p < .05$. No other effect was significant. Tukey's HSD follow-up tests among the four conditions showed that the estimates in the facilitory sufficient condition were significantly different from

the estimates in the facilitory necessary condition, and from the other two inhibitory conditions. No other significant difference was found. For the positive contingencies, the results concurred with the findings from the first two experiments: the sufficient and necessary causes were judged differently. Again, in this experiment, the judgment biases were observed both in the sufficient ($t(47) = -6.29$, $p < .01$) and necessary conditions ($t(47) = 4.63$, $p < .01$).

In the negative games, however, the inhibitory sufficient and the inhibitory necessary conditions were not perceived as significantly different. While the statistics suggest that people might judge the inhibitory sufficient condition and the necessary condition as the same, it is also possible that this kind of contrast was masked in negative contingencies. As shown in Panel A of Figure 4, people's judgments were not negative "enough" relative to the actual contingencies in both inhibitory conditions. This is what causes the significant sign factor in the transformed data analyses. Along the same line of thinking, paired t-tests between the estimates and the normatively negative contingencies showed significant differences were found for both of the inhibitory sufficient condition ($t(47) = 5.28$, $p < .01$) and the inhibitory necessary condition ($t(47) = 3.56$, $p < .01$). However, the differences are in the same direction: they were both less negative than the real contingency. This reluctance to consider the paint as making the tank explode more might be induced by a cognitive schema toward considering anything that is called "camouflage" to be a protecting agent. In addition, the literature on hypothesis testing suggests that people generally do not seek out disconfirming information. They appear to have a distinct bias to search for positive instances of a hypothesis, and they fail to search for negative instances that might disconfirm a hypothesis (Einhorn & Hogarth, 1978; Mynatt, Doherty, & Tweney, 1977; 1978; Snyder & Swann, 1978; Wason, 1968; —although see Evans (1989) for a different interpretation).

The underestimation in negative contingencies may also be relevant to feature-positive effect. Originally, this effect was reported in pigeon discrimination studies by Jenkins and Sainsbury (1969, 1970), and then, observed with human subjects (e.g., Bitgood, Segrave, & Jenkins, 1976; Newman, Wolff, & Hearst, 1980; Norton, Muldrew, & Strub, 1971; Sainsbury, 1971; 1973). For example, Newman et al. (1980) found that when the presence of a special feature was connected with positive meanings (such as a group of cards named "Good"), subjects learned to used it to discriminate stimuli faster than when it was connected with negative meanings (such as a group of cards named "Not Good"). In the negative contingency games, the camouflage made the tank explode more in the minefield. Thus, the camouflage became a negative feature and, accordingly, the subjects might learn the relationship between the camouflage and the outcome more slowly. Other studies also showed that people seem reluctant to judge negative contingencies or to process negative information (See Baker & Mercier, 1989; Jenkins & Ward, 1965; Ward & Jenkins, 1965). Thus, if the subjects were not sensitive enough for the negative contingencies, judgment differences between the sufficient condition and the necessary condition might be too small to show statistical significance at the power level used in testing positive contingencies.

Initial tests of the three theoretical models showed unexpected results. Correlations between the mean estimates and the theoretical predictions were very high for all models. This is because all of them explain the subjects' judgment difference between the positive and the negative contingencies very well. When the data were transformed by inverting the sign of the judgments for all the negative contingency games, the fit of some of the models changed markedly. The ΔP rule showed a large standard deviation (26) of the difference from the judgments, and a low correlation ($r(6) = .35$, n.s.). The adjusted ΔP rule showed an even worse

pattern, SD_{diff} = 27, and $r(6)$ = .27, n.s. For the negative contingencies, it not only accounted for the subjects' estimates inaccurately, but even produced the calculated values in the opposite direction. The adjusted ΔP rule expected that the participants would judge the inhibitory sufficient contingencies as less negative than the inhibitory necessary ones, whereas the participants judged them more negative. The standard simulator with the same parameters as used in Experiment 2 was performed the simulations for the Rescorla-Wagner model, this model provided a much better fit than these rules, albeit less than perfect. The standard deviation of the differences was much lower (8), and the correlation was high ($r(6) = .98, p < .01$). Figure 4 shows the calculated values form the adjusted ΔP rule (Panel C) and from the Rescorla-Wagner model (Panel D).

In an attempt to salvage the adjusted ΔP rule, we reasoned that since the subjects showed a reverse judgment pattern in the negative contingencies, the weights should correspondingly be switched to Cell A and Cell C in the contingency tables. Along with this switch, we also evaluated new values for the weights.

$$Adjusted2\ \Delta P = 1.2A/(1.2A+B) - 1.2C/(1.2C+D) \qquad (12)$$

The weights used in this formula were arrived at by comparing several values and calculating the least square difference from the judgments. Figure 5 (right hand panel) shows these calculated values with the subjects' mean estimates in these negative contingency games (left hand panel). The estimates were less negative than the calculated values, yet the pattern of relations among the means fit the average judgments well. For the four negative contingency games, the correlation between the calculated values and the

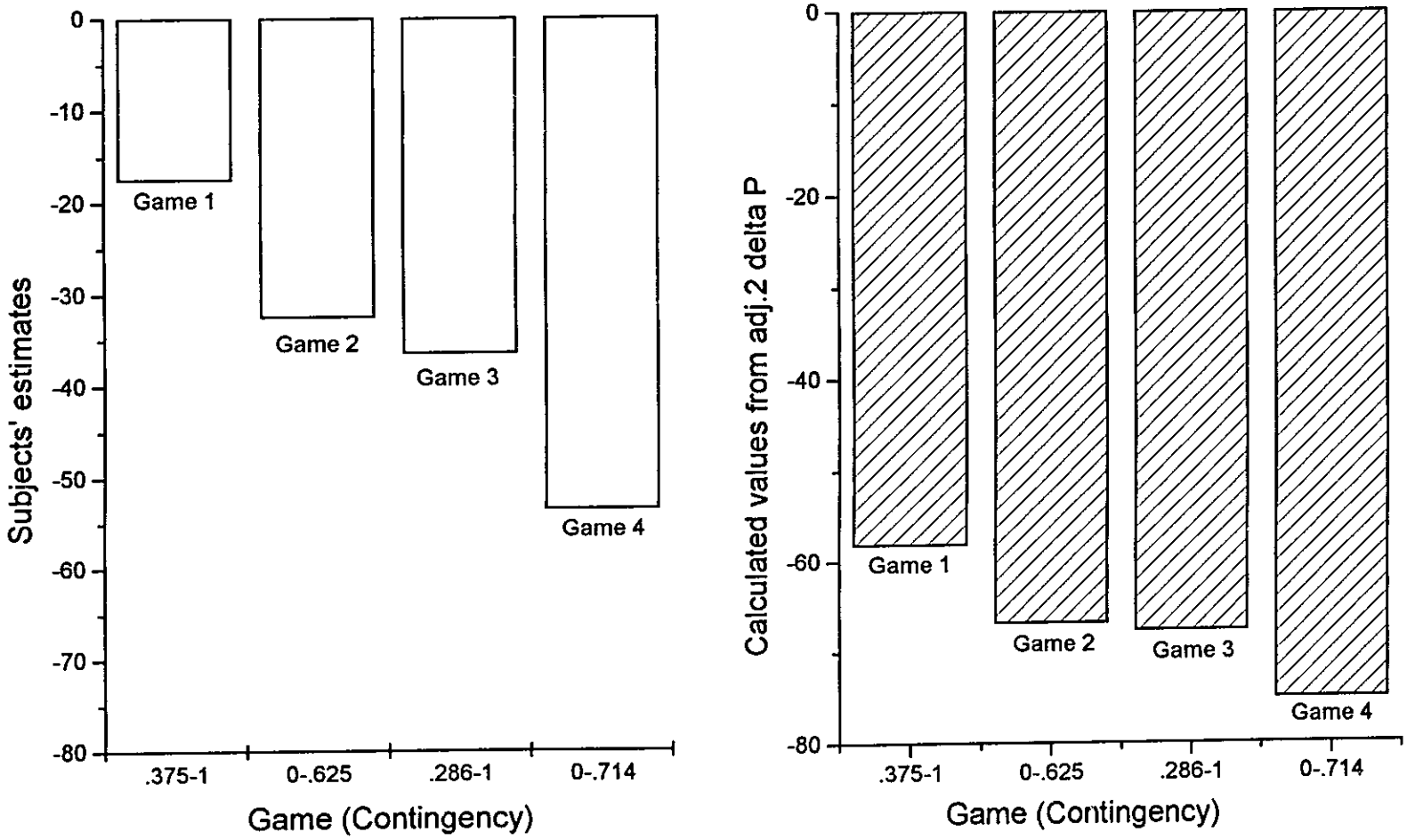

Figure 5. Subjects' mean estimates on negative contingency games in Experiment 3 and the calculated adjusted2 delta P values.

judgments was high, $r(3) = .99$, $p < .01$. However, the calculated values showed a very large SD_{diff} (33). The fact that the calculated values are far lower than the corresponding judgments causes this large standard deviation of the difference, yet, this does not reduce the correlation because the Pearson formula is insensitive to a constant difference. Another drawback of the adjusted2 ΔP is that it does not account for the positive estimates well. It provides a reverse judgment pattern between the facilitory sufficient and the facilitory necessary conditions.

Considering both the positive and negative contingencies, the Rescorla-Wagner model provided the best account for the estimates. With the same parameters as in the previous two experiments, it fits the judgments much better than the two other models. The ΔP rule could not really explain the judgments in these games. The adjusted ΔP rule (Equation 11), with weights on cell B and cell D, successfully explained the positive judgments; however, it could not account for the negative ones. To obtain a better explanation we have to weight different cells in the formula, and combine these two adjusted ΔP rules. Nevertheless, even when these two adjusted ΔP rules were used together, Equation 11 for the positive contingencies, Equation 12 for the negative ones, the correlation between the calculated values and the judgments improved ($r(6) = .62$), but remained non significant and associated with a large standard deviation of the difference (23). In comparing these SD_{diff} ($F(2, 14) = 5.82$, $MSE = 137767$, and $p < .02$), only the Rescorla-Wagner model had an error level significantly lower than ΔP (Tukey). It was also marginally lower ($p < .06$) than the combination of the two adjusted ΔP.

Multiple regression showed that when ΔP was forced to enter the formula first, its contribution was not significant and it only accounted for 12 percent of the variance of the estimates. The addition of the Rescorla-Wagner model in the equation accounted for an additional

85 percent of the variance ($p < .01$). The two adjusted ΔP rules combined together only accounted for an added 27 percent of the variance and was not significant.

Experiment 4

The first three experiments found that both the adjusted ΔP rule and the Rescorla-Wagner model can explain the positive contingency judgments very well. However, the goodness of fit of the adjusted ΔP rule and the Rescorla-Wagner model was accomplished by adjusting their parameters post-hoc. The final adoption of the parameters was done by selecting values which best fitted the pre-recorded judgments. Thus, we did not really test the predictive power of these models. In Experiment 4 the predictive power of the adjusted ΔP rule and the Rescorla-Wagner model was compared with that of the ΔP rule.

Experiment 4 included four tank games as shown in Table 8. These games are all positive contingencies. They were created according to the predicted values of the different models. The ΔP and the adjusted ΔP rule (Equation 11) predict different contingencies for these games. For example, ΔP predicts that the participants should make the same judgments in Game 1 and Game 3, as well as in Game 2 and Game 4. However, the adjusted ΔP rule, with weights on cell B and cell D, predicts different judgments within the two pairs of games. The Rescorla-Wagner model, with the same parameters used in Experiments 2 and 3, predicts judgments similar to the adjusted ΔP rule, but with larger differences within the two pairs of games. It predicts mean judgments lower than the adjusted ΔP in Games 3 and 4. A counter-intuitive prediction made by the adjusted ΔP rule and the Rescorla-Wagner model is that of no judgment difference between Game 1 and Game 2, while the ΔP values for these games differ by as much as .12. Figure 6 shows the predicted values from the ΔP rule (Panel A), the adjusted ΔP rule (Panel B), and the Rescorla-Wagner model (Panel C). The order of the panels was reversed from the first three experiments to reflect the predictive nature of this test.

Table 8

Frequencies of Camouflaging Paint and Explosion in the 20 Trial Periods of Each Game in Experiment 4

	SAFE	EXPLODE
PAINT	8	1
NO PAINT	2	9

GAME 1
ΔP=.889-.182=.707
Adj. ΔP=70.7
R-W model=69

	SAFE	EXPLODE
PAINT	3	0
NO PAINT	7	10

GAME 2
ΔP=1-.412=.558
Adj.ΔP= 71.4
R-W model=70

	SAFE	EXPLODE
PAINT	10	4
NO PAINT	0	6

GAME 3
ΔP=.714-0=.714
Adj. ΔP=58.8
R-W model=50

	SAFE	EXPLODE
PAINT	10	7
NO PAINT	0	3

GAME 4
ΔP=.588-0=.588
Adj. ΔP=49.9
R-W model=34

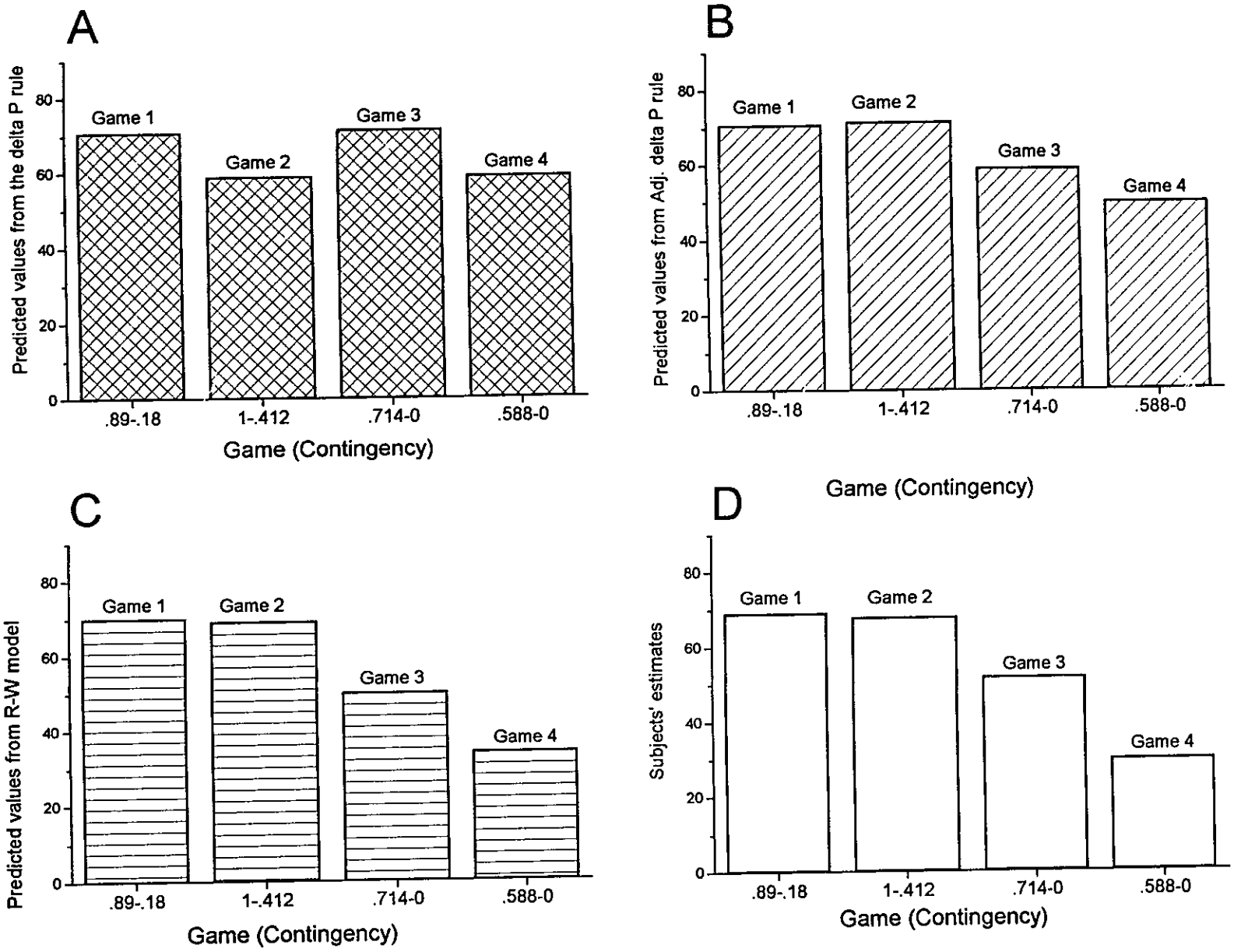

Figure 6. Predicted values from the delta P rule, the adjusted delta P rule, and the Rescorla-Wagner model; and the subjects' mean estimates in Experiment 4.

Method

Participants and apparatus. The participants were 24 undergraduates from the University of Ottawa. When these participants were recruited, they were informed that one prize of $50.00 would be drawn randomly and given to one of them for their participation. The apparatus and the stimuli described in Experiment 1 were used.

Procedure. The same instructions described in Experiment 1 were used again. Counterbalancing and randomization were the same as in the previous three experiments.

Results and Discussion

The subjects' mean guess of the camouflage was 21.87. Figure 6 shows the mean final judgments in each game (Panel D) alongside the predicted values. Different judgments were made in different games. Participants judged the paint efficacy as similar in Game 1 and Game 2. However, they made different judgments between Game 1 and Game 3, and between Game 2 and Game 4.

A 4 X 2 within-subject analysis of variance, for different games and for the judgments after 20 trials and 40 trials, was carried out. There was a significant main effect of games ($F(3, 69) = 11.96$, $MSE = 1344$. $p < .01$). There was no significant difference for the judgments made after 20 or 40 trials and no significant interaction was found. Tukey's HSD post hoc tests showed that the judgments in Game 4 were significantly different from that in any of the other three games. The judgment difference between Game 1 and Game 3 was marginally significant ($p < .08$) and so was the difference between Game 2 and Game 3 ($p < .11$). Since the comparisons between Game 3 and either Game 1 or Game 2 were planned, tests of priori contrasts were

conducted on these pairs of games. These tests showed a significant difference between Game 1 and Game 3, $F(1, 23) = 5.75$, $MSE = 1218$, $p < .05$; and between Game 2 and Game 3, $F(1, 23) = 5.04$, $MSE = 1546$, $p < .05$. The difference between Game 1 and Game 2 was not significant, $F(1, 23) = .027$.

The statistical analyses showed that the adjusted ΔP rule and the Rescorla-Wagner model predicted the subjects' judgments very well in these games, while the ΔP rule did not. The correlations between the mean judgments and the predicted values confirmed this assertion. For the ΔP rule, the correlation coefficient between the mean judgments and the predicted values was not significant, $r(2) = .35$. The correlations for the adjusted ΔP rule ($r(2) = .987$, $p < .05$) and the Rescorla-Wagner model ($r(2) = .993$, $p < .01$) were both very high. Multiple regression analyses also showed similar results. When ΔP entered the formula first, it could account for 12 percent of the variance of the mean judgments. Adding the adjusted ΔP rule could account for 88 percent more variance, while the Rescorla-Wagner model accounted for an added 87 percent. On the other hand, the Rescorla-Wagner model has the advantage that its standard deviation of the differences was very small (2), while the deviation for the adjusted ΔP (11) and the ΔP (18) were quite large. Unfortunately no significant difference ($F(2, 6) = 2.96$, $MSE = 36261$) could be detected among these error measurements due to the low power of the test involving only four data points per model.

From these results, it may be concluded that the Rescorla-Wagner model can predict the correct pattern of judgments and do so rather accurately. The adjusted ΔP rule predicted the judgment pattern well, but was not as accurate the Rescorla-Wagner model. The ΔP rule was the least accurate and did not predict the pattern of means at all.

General Discussion

The four experiments reported above extend our understanding of covariation detection in single contingency task. These experiments included a series of positive contingencies in the range .59 to .77, while the negative contingencies studied were -.625 and -.714. These ranges were chosen to allow for multiple frequency tables with each ΔP value while keeping the outcome density (10 per period) constant across all games in the four experiments. Within this narrow range of contingencies the pattern of results is related to the causal status of the covarying events and has implications for the evaluation of different computational models.

Individual Data versus Group Data

In the four experiments, the data were reported and analyzed at the level of group means. The evaluation of the different models was also based on the group data. Thus, these models were tested as descriptive ones to explain the judgments of a group of people and predict mean judgments of the population. As discussed in the introduction section, people's judgments can also be described and analyzed at the individual level, and the relevant models can be evaluated at this level as well. The very best model would be one that could successfully explain both the mean judgments of a group of subjects and the individual judgments. However, we have a long way from understanding human judgment processes completely, down to the level of individual differences. Thus, this thesis relies on group data, considering this strategy as optimal at this point of time.

Individual differences in contingency judgments have often been observed (e.g., Kao & Wasserman, 1993; Shaklee & Min, 1986; Wasserman et al., 1993). In the present four

experiments, subjects were also observed to make very different judgments for the same game. For example, in Experiment 3, where Game 2 was a negative contingency ($\Delta P = -.625$) presented under an inhibitory sufficient condition, most subjects gave negative values to the camouflage paint; however, four subjects judged it as 0 and another four subjects gave positive numbers. For the positive contingencies, individuals also showed great judgment differences. For instance, in Experiment 4, where the ΔP value in Game 3 equals .714, the subjects' estimates in this game ranged from -100 to +90.

As discussed in the introduction section, in some previous off-line studies, researchers have supposed that the different individuals use different rules. A rule analytic technique was used to discriminate the different rule users (e.g., Shaklee & Mims, 1986). The discrimination was based on two assumptions: 1). each individual uses one of the postulated rules to make his or her judgments and 2) each individual uses only one rule across all conditions or problems. In reality, these methods can only correctly identify the subjects who have solved a certain category of problems. There is no assurance that these subjects have used one rule and one rule only. It is also possible that some subjects may have used some rules that were not included in the hypothesized list, or even that some or all of them used no rules at all, but that their performance on the judgments was identical to the supposed single rule users. This last argument was supported by some studies. For instance, in experiments of Kao and Wasserman (1993) and Shaklee & Mims (1986), several subjects could not be categorized as using any specific rule. In the present experiments, when a best parameter search was performed on the weighted ΔD rule (Equation 5) to differentiate individual rule users, only a few subjects fell into Kao and Wasserman's practical criterion of "cell A versus B" category. In addition, for these subjects, the

calculated values from the weighted ΔD were very different from the individual's real judgment. In Experiment 4 for example, one of these subjects showed a standard deviation of the difference from the judgment as large as 42. It is doubtful then that these subjects really used the cell A versus cell B rule to make their judgments. To extend the limitations for their practical criteria, only a few more subjects could be identified as using the information in the four cells (ΔD rule users) or as using cell A versus B. The remaining large number of subjects did not belong to any those postulated rule users. Out of the total 96 subjects, 8 subjects could be classified as using cells A, B, and C only; 5 appeared to use cells B, C, and D only, 18 seemed to use cells A, B, and D; and 10 only would have used cells B and D. Together with the previous studies, these results suggest that in making their judgments, some subjects may use rules that have not been proposed, some might not use any rule at all, or some might make their judgments based on using only a portion of the information presented to them. It is also reasonable to suppose that some subjects may use different rules in different problems.

The probabilistic contrast model (Cheng & Novick, 1992) and the Rescorla-Wagner model (1972) address the issue of individual differences unlikely from the rule analytic strategy. Both of them proposed that every subject has only one computation: the probabilistic contrast or the increment/decrement of associative strength. According to the probabilistic contrast model, individuals all compute probabilistic contrasts but they may differ in what information they select for computation. For example, in a double contingency situation where some of the cues may never appear alone (e.g., tank camouflage and spotter plane or two symptoms for one disease), there is ambiguity as to which pieces of evidence should enter in the probabilistic contrast formula. According to Melz et al. (1993), some subjects may choose to compute a probabilistic

contrast for cue 2 conditional on the *presence* of cue 1 whereas others might prefer to conditionalize on the *absence* of cue 1 (if the relevant information is available). This will produce two different focal sets on which to base the contrast with corresponding individual differences. Unfortunately, Melz et al. have not stated how such individual choices could be predicted. Note, however, that in single contingencies such as in the current experiments, there is no ambiguity as to which data should be used and all participants should rely on the universal set of data. Thus the probabilistic contrast model is silent about individual differences for single contingencies.

In the Rescorla-Wagner model, associative strength is influenced by the order in which the information is presented. Thus, one source of individual judgment differences may be due to the different random orders of trials that they each experienced. Another potential source of individual differences lies in the possibility of different levels of stimulus salience which could be reflected in α and β values for each participant. Again, however, the model has no decisional mechanism for setting such values.

Although the data analyses and the theoretical simulations presented in the core of the thesis rest on group means, an attempt was made to incorporate individual differences. A search for the best individual weights to adjusted ΔP was preformed using the HJ method. Even when each subject was assigned his/her own adjustment, the SD_{diff} for the adjusted ΔP calculation remained larger than when the corresponding calculation was carried out at the group level (by Equation 11).

For the Rescorla-Wagner model, simulations were carried out with: 1) best values (obtained from Experiment 2) for parameters α and β but incorporating in the iterative calculations the exact sequence of trials experienced by each subject; 2) best individual values

for α and β (Experiments 2 and 3 only), while keeping the different sequences of the trials for different subjects as they were experienced. Neither simulation produced a fit as good as the simulation performed on group means in terms of SD_{diff}.

All these analyses suggest that we have not found the best model to deal with the individual judgments yet. It was possible that some subjects used one rule, or followed one model through all judgments, but the others might use different rules or weight different information differently under different conditions. The issue remains open. Thus, at this point of time, the evaluation of the various competing models must considered valid at a descriptive level only.

Sufficiency versus Necessity

Sufficient and necessary conditions were used in all four experiments. Participants estimated the facilitory sufficient and necessary factors differentially. They overestimated the efficacy of the sufficient factor, and underestimated the necessary one. The judgments were biased in both conditions. For the inhibitory sufficient and necessary conditions, it was found that judgments mirrored the pattern observed for positive contingencies, yet the differences did not reach statistical significance, possibly because of an overall insensitivity to negative contingencies.

In the social attribution and developmental studies, the status of sufficient and necessary conditions remains controversial. Having different purposes, these studies employed different research paradigms but they agreed that a factor leads to the strongest perception of causation when it is both necessary and sufficient for an outcome. If a factor is neither necessary nor sufficient, then it elicits the weakest causal judgment. The relative impact of a factor when it is

either necessary or sufficient, but not both, is less clear. Many previous studies did not differentiate them, suggesting that they both have an intermediate impact (e.g., Bindra, Clarke, & Shultz, 1980; McGraw, 1987; O'Brien & Davidson, 1989; Pennington & Hastie, 1983; Surber, 1981). However, a few studies found that sufficiency is more important than necessity (e.g., Schustack & Sternberg, 1981) while a few others found the opposite (e.g., Siegler, 1976). These differences are usually explained as the influence of question difficulty or task differences (see McGraw, 1987). However in the current experiments, the sufficiency/necessity differences were observed within task and thus did not appear to be artificial.

In the four experiments reported above, the concepts of necessity and sufficiency were combined with contingency, and tested on-line. The results showed a direct relation between these concepts and the *direction* of the observed biases. Overall however, as real as the necessity/sufficiency distinction may be at philosophical level, empirically, it can only help us to determine the direction of expected biases. To obtain quantitative estimates of the biases, we need to turn to computational models of contingency judgment.

Evaluation of the ΔP Rule

When the contingency estimates obtained in Experiments 1-4 were within the same condition (sufficient/necessary), the subjects' judgments were consistent with ΔP: as ΔP increased the subjects' judgments increased as well. This aspect of the results coincides with previous reports (e.g., Alloy & Abramson, 1979; Allan & Jenkins, 1980, 1983; Baker et. al., 1989; Dickinson et. al., 1984; Shanks, 1987; Wasserman et. al., 1983, 1993). However, when a constant contingency is examined across the necessary condition, the sufficient condition, and the neither

necessary nor sufficient conditions, the judgments are no longer consistent with ΔP. For example, in Experiment 4, the mean judgments were almost the same for Games 1 (68.96) and 2 (67.71), but the ΔP rule predicted different values (70.7 for Game 1 and 58.8 for Game 2). Here the camouflage paint was a sufficient factor in Game 2, but it was not in Game 1. Additionally, whereas the ΔP rule expected almost no judgment difference between Games 1 (70.7) and 3 (71.4), the subjects' estimates actually differed significantly (68.96 for Game 1 and 51.67 for Game 2). The paint was presented as a necessary factor in Game 3, but not in Game 1. The same situation happened in Games 2 and 4. The judgments are affected by how these contingencies are formed in the corresponding contingency table. Thus, one of the main findings from the present experiments is that the ΔP rule cannot properly account for mean judgments based on group data in single contingency tasks.

This finding is in direct contradiction with the probabilistic contrast model proposed by Cheng and Novick (1990a, 1992). Cheng and Novick (1990b) originally reviewed the literature in the social/cognitive research tradition and concluded that there was no solid evidence of bias when people were asked to judge the relation in one pair of events (single contingency task). They did find attributive biases in multiple relations and developed their model mostly to account for discounting effects (Cheng & Novick, 1990a; 1992; Melz et al., 1993). However, the model clearly stated that for single contingencies, people's judgments should be made, normatively, on the basis of the ΔP rule (Cheng & Novick, 1992, p.367). If people's judgments were considered qualitatively only, it might be said that the model is correct. In the current four experiments, when the ΔP value, or the probabilistic contrast between the two probabilities was positive, the subjects judged the paint as a facilitory factor to protect the tank; when the value was negative,

an inhibitory judgment was obtained. Quantitatively, however, the probabilistic contrasts cannot account for the judgment changes across different games.

To account for the present data while retaining the probabilistic contrast as a valid computational model, one would have to suppose that the judgment biases observed come from an erroneous sampling of the data and not from a computational error. In a study relevant to the sufficiency/necessity issue, Wasserman et al. (1993) had subjects reporting to what extent their tapping a telegraph key increased or decreased the probability of a small light flashing. Each contingency task lasted about 1 minute, and this period of time was divided into 60 one-second sampling intervals. During each interval, if the subject tapped the key at least once, the light would flash at the end of the interval, then another sampling interval began. If there was no outcome in the sampling interval (no light flashed), then another sampling interval automatically began. Although they were not specifically testing for sufficiency and necessity, Wasserman and colleagues reported a pattern of results which appeared to be reverse of the current one for the positive contingencies. However, the key tapping tasks differ from the tank game in at least two aspects. First, the subjects were active; they could decide how many times they tapped the key in each 60-second task. For every single contingency, differences in tapping frequency could be found among the subjects. Thus, strictly speaking, the subjects were not shown the same stimuli, or the same contingency respectively. These sampling differences may have influenced the judgment patterns of the subjects. Second, the sampling interval was very short, and there was no signal to tell the subjects it was an interval. When the researchers calculated the ΔP values for each contingency, they must have used 60 (the number of total sampling intervals in each contingency) as the total number of the four cells in the contingency table. But their subjects

might have counted these frequencies, especially cell D, differently (if indeed what people do is counting frequencies). Cell D is the frequency of no-tap/no-light. Since each sampling interval was arbitrarily set to 1 second, if the subjects did not tap for a while, and there was no light flashing, they might only have added one to cell D, even if there might be two or more seconds passing. Meanwhile, the researchers added a larger number to cell D. If it is supposed that the subjects applied the ΔP rule to this data sample, we can explain the reversal of their subjects' judgment pattern in the necessary and sufficient conditions of positive contingencies. In the sufficient condition of the positive contingencies, the cell B equalled 0 and the value of P(O|A) was 1, an extreme value. The researchers and the subjects might not have differed in this calculation. For P(O|~A), since the subjects might have used a smaller number for cell D, the formula (P(O|~A) = C/(C+D)) would yield a greater probability for the subjects than for the researchers. Consequently. the total ΔP value appeared smaller. However, for the necessary condition, since cell C equalled 0, the smaller frequency of cell D counted by the subjects would not have influenced the ΔP values. This analysis is supported by their Experiments 2 and 3. In these two experiments, the subjects were also asked to estimate P(O|A) and P(O|~A) respectively. Their subjects estimated most of the P(O|~A) values higher than the real ones.

Can a similar sampling bias account for the present data? This is unlikely because all trial types, even cell D, where clearly identified in our experiments. The participants were shown clear instances of the presence and absence of the explosion, with or without the paint since the passage of the tank on the screen marked each trial clearly. Also, between two trials, the subjects had to make a guess about the explosion of the tank. So, they had enough time to distinguish each trial from the previous one and the subsequent one. Thus, the task itself did not tend to bias

data sampling. It is still possible, of course, that the participants disregarded some of the information presented to them. Indeed, there is some evidence that people do not regard the information contained in the four cells of contingency tables as equally important (Wasserman, Dorner, & Kao, 1990). Unfortunately however, the probabilistic contrast model of Cheng and Novick does not address this issue. In any case, it would be more parsimonious to have a model that could account for all biases. The Rescorla-Wagner model is just such a candidate.

Evaluation of the Rescorla-Wagner Model

The explanatory power of the Rescorla-Wagner model has already been applied to the overshadowing effect in contingency judgments (e.g., Baker, et al., 1993; Shanks, 1993; Vallée-Tourangeau et al., 1994). For single contingencies, it has the additional strength of being able to deal with any contingency table, no matter how many cells are empty. For example, a given contingency may contain trials in cell A and B only (e.g., partial reinforcement). People can easily make judgments based on these trials; however, the ΔP rule cannot generate a value since, when cells C and D equal 0, $P(O|\sim A)$ is mathematically indeterminate. To let a probabilistic contrast model handle such cases, one would have to assume that the cognitive system uses a single conditional probability when the full ΔP rule is unapplicable. The Rescorla-Wagner model does not require any exception to its central rule.

In this model, different values can be used as the starting point of VA (in Equation 8). Such a value should reflect any prior knowledge about a given factor. In the present experiments, the average guess was used as the starting value for $V_{camouflage_0}$. However, the final associative strength generated by these simulations did not differ from those obtained by relying on an initial

value of zero.

Based on the data obtained, the Rescorla-Wagner model provides an impressive explanation and prediction of people's mean judgments. Using only one set of parameters, the model accounted for the subjects' judgments rather well in all four experiments. The model fit the data both in terms of reproducing the patterns of means (high correlation) and in terms of minimizing error (standard deviation of the differences from the norm). Multiple regression analyses further supported this conclusion. In terms of minimizing error, the hypothesis tests conducted on the SD_{diff} of individual experiments often failed to reach significance. These tests had little power because of the small sample of differences. However, an overall ANOVA pooling all four experiments ($F(2, 54) = 6.76$, $MSE = 63635$, $p < .05$) revealed that the mean error for the Rescorla-Wagner model (with the same parameters over the four experiments) was significantly smaller than that of ΔP (Tukey). No other comparisons were significant among the Rescorla-Wagner model, the adjusted ΔP rule (combined adjusted ΔP was used for Experiment 3) and the ΔP rule.

This is not to say that the associative model is without weaknesses. One weak point is that we have no prior criterion for selecting values of the salience parameters (α, β). These have to be derived empirically from some initial data set obtained from the same participants in the same context. Another important aspect of the associative simulations, as discussed in the section on individual versus group data, is that using the best fit can only be obtained for the mean judgments of a group of subjects. To be useful at the individual data level, additional specifications are required. Keeping these limitations in mind, the Rescorla-Wagner model remains a very good tool to describe and predict people's judgments of covariations in both single

and double contingency tasks.

Evaluation of the Adjusted ΔP Rule

The adjusted ΔP rule showed a much better fit to the data than the simple ΔP did. Its correlation with the mean judgments was often the same as that of the Rescorla-Wagner model.

This adjusted rule keeps the basic ΔP formula for all mathematical operations on the corresponding cells. The difference between this formula and ΔP is that a weight is added to some cells. For example, in Experiments 1, 2, and 4, a weight of 1.75 was added to cells B and D (Equation 11). Considering the mean judgments in these experiments, the calculated values from the simple ΔP are not large enough in the sufficient conditions and not small enough in the necessary conditions. In the sufficient condition, P(O|A) has already reached the maximum value of 1. The only way to produce a higher calculated contingency based on the ΔP method is to decrease P(O|~A). Similarly, in the necessary conditions, P(O|~A) is at its minimum of 0, since cell C is equal to 0. The value of P(O|~A) can not be increased. So, P(O|A) must be decreased. The adjusted ΔP rule (Equation 11) does exactly what we need to modify the two conditional probabilities. In the sufficient condition, the weight for cell B has no effect since cell B equals 0, but the weight for cell D decreases the value of P(O|~A). In the necessary condition, the weight for cell D has no effect since cell C is equal to 0, but the weight for cell B makes the value of P(O|A) decrease.

In Experiment 3, the subjects' estimates of the negative contingencies could not be modeled using the same weights on the same cells. As shown in Equation 12, Cells A and C had to be adjusted instead of B and D in order to obtain a judgment pattern matching the data. Thus,

two kinds of weights had to be combined to account for all the experiments. This is less parsimonious than the Rescorla-Wagner model. Moreover, as discussed in Experiment 3, even when the weights are allowed to shift, the predictions of the adjusted ΔP remained associated with a rather large SD_{diff} and lower correlation than the Rescorla-Wagner model.

Adjusting ΔP by weighting the cells of the contingency table is conceptually analogous to adjusting the salience parameters α and β in the Rescorla-Wagner model. Since it has also been shown that the Rescorla-Wagner model matches the ΔP values at asymptote (Melz et al., 1993; Shanks, 1993; Wasserman et al., 1993), it could be argued that the empirically derived weights that we used in adjusted ΔP rule were not the best and that another set of weights imitating the α and β derived for the associative model would make a better adjustment for ΔP. To test this possibility, cells A and B were weighted by α but not cells C and D, cells A and C were weighted by β value for the outcome, and cells B and D by the β value for no-outcome. These parametric values made the Rescorla-Wagner model successful in Experiment 3. However, this alternative adjustment to ΔP did not fit the data very well.

Using the HJ method (Kao & Wasserman, 1993) to adjust the ΔP rule (Equation 4), a search was done for the best weight parameters to explain the mean judgments obtained in Experiment 3. The best weights for ΔP were based on the least squared difference between the mean estimates and the calculated values of *all* games in Experiment 3. Again, no set of weights could provide a good fit. For instance, using the best weights, the calculated values predicted almost no difference at all between the sufficient and the necessary conditions (e.g., Game 5, ΔP = .625, necessary condition, calculated value = 80.27; Game 6, ΔP = .625, sufficient condition, calculated value = 80.64).

As a last effort to see if some kind of weighting scheme could account for the data obtained in Experiment 3, the HJ method was applied to Kao and Wasserman's (1993) non-normative information formula (Equation 5) for the mean estimates. This time, with different weights on different cells, rather good fit was obtained for Experiment 3. The SD_{diff} was only 3. The same equation was also found to fit the mean judgments well in the other three experiments. However, the best weights for these experiments were not the same as those for Experiment 3. For all the experiments, cell A had the largest weight, followed by cell B. Cell C was fairly constant with a low weight ($W_c < .05$). For cell D, different results were obtained for different experiments: the weight was very small in Experiments 2 and 4 ($W_D < .02$), and became larger and ignorable in Experiments 1 (.12) and 3 (.19). Thus, in spite of its potential, the averaging model of Kao and Wasserman is not as parsimonious as the associative model of Rescorla-Wagner since the averaging method required a different set of weights for each experiment instead of one set of parameters for all experiments as the associative model does.

We have already mentioned that an associative model based on the Rescorla-Wagner equations is not without its weakness. Nevertheless, in the context of the four experiments presented here, and taking into account its successes in many previous publications, it seems fair to conclude that this model remains the best descriptive theory to date for explaining human contingency judgments.

References

Adi, H., Karplus, R., Lawson, A. and Pulos, S. (1978). Intellectual development beyond elementary school: VI, Correlational reasoning. *School Science and Mathematics, 78*, 675-683.

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

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canad J Psychol/Rev canad Psychol, 34*, 1-11.

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

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108*, 441-485.

Anderson, J.R. (1990a). *The Adaptive Character of Thought*. Hillsdale: L. Erlbaum Associates.

Anderson, J.R. (1990b). *Cognitive Psychology and Its Implications*. 3rd ed. New York: W.H. Freeman and Co.

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

Arkes, H. R., & Rothbart, M. (1985). Memory, retrieval, and contingency judgments. *Journal of Personality and Social Psychology, 49*, 598-606.

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

Baker, A. G., & Mercier, P. (1989). Attention, retrospective processing and cognitive representations. In S. B. Klein & R. R. Mowrer (Eds.), *Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theory*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Bindra, D., Clarke, K. A., & Shultz, R. (1980). Understanding predictive relations of necessity and sufficiency in formally equivalent "causal" and "logical" problems. *Journal of Experimental Psychology: General, 4,* 422-443.

Bitgood, S. C., Segrave, K., & Jenkins, H. M. (1976). Verbal feedback and the feature-positive effect in children. *Journal of Experimental Child Psychology, 21,* 249-255.

Blashfield, R. K., & Aldenderfer, M. S. (1988). The methods and problems of cluster analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology*. New York: Plenum Press.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory* (Vol. 1, pp. 187-215). Hillsdale, NJ: Erlbaum.

Capon, N. and Kuhn, D. (1979). Logical reasoning in the supermarket: Adult female's use of a proportional reasoning strategy in an everyday context. *Developmental Psychology, 15,* 450-452.

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

Cheng, P. W. & Nisbett, R. E. (1993). Pragmatic constraints on causal deduction. In R. E. Nisbett (ed.), *Rules for Reasoning (pp. 207-227).* Hillsdale, New Jersey: Lawrence Erlbaum Association.

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

Cheng, P. W., & Novick, L. R. (1990b). Where is the bias in causal attribution? in K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & Erdos (Eds.), *Lines of thinking: Reflection on the psychology of thought (Vol. I)*. New York: John Wiley & Sons Ltd.

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

Cheng, Y., & Mercier, P. (1990). Permanent blocking in human contingency judgments. Poster presentation on *Annual Convention of CPA, 1990, Ottawa.*

Cronbach, J. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50,* 456-473.

Dickinson, A. (1980). *Contemporary animal learning theory.* New York: Cambridge University Press.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A*, 29-50.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review, 92*, 433-461.

Ericsson, K. and Simon, H. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hilssdale: Lawrence Erlbaum Associates.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

Goldberg, L. (1968). Simple models or simple processes? *American Psychologist, 23*, 483-496.

Hempel, C. (1964). *Aspects of scientific explanation.* New York: Collier Macmillan.

Hewstone, M. (1983). Attribution theory and common-sense explanations: An introductory overview. In: M. Hewstone (Ed.), *Attribution theory: Social and functional extensions.* Oxford, England: Basil Blackwell.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic.

Jenkins, H. M., & Sainsbury, R. S. (1969). The development of stimulus control through differential reinforcement. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental issues in associative learning.* Halifax, Nova Scotia, Canada: Delhousie University Press.

Jenkins, H. M., & Sainsbury, R. S. (1970). Discrimination learning with the distinctive feature on positive or negative trials. In D. Mostofsky (Ed.), *Attention: Contemporary theory and analysis.* New York: Appleton-Century-Crofts, 1970.

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

Jennings, D. L., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior.* New York: Appleton-Century-Crofts.

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

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (ed.), *Nebraska Symposium on Motivation (Vol. 15).* Lincoln: University of Nebraska Press.

Kelley, H. H. (1972). *Causal schemata and the attribution.* New York: General Learning Press.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28,* 107-128.

Mackintosh, N. J. (1971). An analysis of overshadowing and blocking. *Quarterly Journal of Experimental Psychology, 23,* 118-125.

Mackintosh, N. J. (1974). *The psychology of animal learning.* London: Academic Press.

McGraw, K. M. (1987). Conditions for assigning blame: The impact of necessity and sufficiency. *British Journal of Social Psychology, 26,* 109-117.

Melz, E. R., Cheng, P. W., Hoyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner learning rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1398-1410.

Mercier, P. (in press). Computer simulations of the Rescorla-Wagner and the Pearce-Hall models in conditioning and contingency judgement. *Behavior Research Methods, Instruments, and Computers.*

Miles, C. G., & Jenkins, H. M. (1973). Overshadowing in operant conditioning as a function of discriminability. *Learning and Motivation, 4,* 11-27.

Mill, J. S. (1843). *A system of logic.* London: Oxford University Press.

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

Miller, R. R. and Matzel, L. D. (1989). Contingency and relative associative strength. In: S.B. Klein and R.R. Mowrer (Eds.), *Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theory,* (pp. 61-80). Hillsdale, New Jersey: Lawrence Erlbaum Associates,Inc.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 24,* 326-329.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology, 30,* 85-96.

Newman, J., Wolff, W. T., & Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 630-650.

Newell, A. (1980). Reasoning, problem-solving, and decision processes: The problem space as a fundamental category. In: R. Nickerson (Ed.), *Attention and performance VIII.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Norton, G. R., Muldrew, D., & Strub, H. (1971). Feature-positive effect in children. *Psychonomic Science, 23,* 317-318.

O'Brien D.P., & Davidson, G. M. (1989). Evaluation of evidence for sufficiency, for necessity, and for necessity-and-sufficiency. *Quarterly Journal of Experimental Psychology,* 41A, 531-551.

Pennington, N., & Hastie, R. (1983). *Conditions for Inferring Cause.* Paper delivered at the meeting of the Interdisciplinary Conference, Jackson Hole, WY.

Popper, K. (1959). *The logic of scientific discovery.* New York: Basic Book.

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

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

Sainsbury, R. S. (1971). The "feature positive effect" and simultaneous discrimination learning. *Journal of Experimental Child Psychology, 11,* 347-356.

Sainsbury, R. S. (1973). Discrimination learning utilizing positive or negative cues. *Canadian Journal of Psychology, 27,* 46-57.

Schustack, M. W. (1988). Thinking about causality. In R. J. Sternberg & E. E. Smith (Eds.), *The psychology of human thought.* New York: Appleton-Century-Crofts.

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

Seggie, I. (1987). The judgment of covariation between binary variables: Some conditions that influence the process. *Memory & Cognition, 15,* 341-348.

Seligman, L. (1975). *Helplessness: On depression, development. and death.* San Francisco: W. H. Freeman.

Shaklee, H. (1983). Human covariations judgment: Accuracy and strategy. *Learning and Motivation, 14,* 433-448.

Shaklee, H. and Fischhoff, B. (1982). Strategies of information search in causal analysis. *Memory & Cognition, 10,* 520-530.

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

Shaklee, H. and Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development, 52,* 317-325.

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

Shaklee, H. and Mims, M. (1986). Development of rule use in judgment of covariation between events. In: H.R. Arkes and K. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader.* Cambridge: Cambridge University Press.

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

Shaklee, H., & Wasserman, E. A. (1986). Judging interevent contingencies: Being right for the wrong reasons. *Bulletin of the Psychonomic Society, 24,* 91-94.

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

Shanks, D. R. (1985b). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology, 37B,* 1-21.

Shanks, D. R. (1986). Selective attribution and the judgment of causality. *Learning and Motivation, 17,* 311-334.

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

Shanks, D. R. (1990a). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology, 42A,* 209-237.

Shanks, D. R. (1990b). Connectionism and human learning: Critique of Gluck and Bower (1988). *Journal of Experimental Psychology: General, 119,* 101-104.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433-443.

Shanks, D. R. (1993). Associative versus contingency accounts of category learning: Reply to Melz, Cheng, Holyoak, and Waldmann (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1411-1423.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The Psychology of learning and motivation (Vol. 21).* New York: Academic Press.

Shimazaki, T., Tsuda, Y., & Imada, H. (1991). Strategy changes in human contingency judgments as a function of contingency tables. *The Journal* of General Psychology, 118, 349-360.

Shultz, T. R., Butkowsky, I., Pearce, J. W., & Shanfield, H. (1975). Development of schemes for the attribution of multiple psychological causes. *Developmental Psychology, 11,* 502-510.

Shultz, T. R. & Mendelson, R. (1975). The use of covariation as a principle of causal analysis. *Child Development, 46,* 394-399.

Siegler, R. S. (1976). The effects of simple necessity and sufficiency relationships on children's causal inferences. *Child Development, 47,* 1058-1063.

Surber, C. F. (1981). Necessary versus sufficient causal schemata: Attribution for achievement in difficult and easy tasks. *Journal of Experimental Social Psychology, 17,* 569-586.

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

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36,* 1202-1212.

Tang, C.S., & Critelli, J.W. (1990). Depression and judgment of control: Impact of a contingency on accuracy. *Journal of Personality, 58,* 717-727.

Tversky, A., & Kahneman, D. (1974). *Judgment under uncertainty: Heuristic and biases.* Science, 185, 1124-1131.

Tversky, A., & Kahneman, D. (1980). Casual schemas in judgments under uncertainty. In M. Fishbein (ed.), *Progress in social psychology (vol. 1).* Hillsdale, N. J.: Lawrence Erlbaum Associates.

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

Wagner, A. R. (1969). Stimulus validity and stimulus selection in associative learning. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental issues in associative learning.* Halifax: Dalhousie University Press.

Wagner, A. R. (1971). Elementary association. In H. H. Kendler & J. T. Spence (Eds.), *Essays in neobehaviorism: A memorial volume to Kenneth W. Spence.* New York: Appleton-Century-Crofts.

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

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

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20,* 273-281.

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

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning,Memory,and Cognition, 16,* 509-512.

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