uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Mélanie Clément
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Psychology)
GRADE / DEGREE

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

Aspects of Memory Capacity and Confidence in Contingency Judgements

TITRE DE LA THÈSE / TITLE OF THESIS

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

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

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Alain Desrochers

Claudette Fortin

Pierre Gosselin

Catherine Plowright

Gary W. Slater
LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES /
DEAN OF THE FACULTY OF GRADUATE AND POSTDOCORAL STUDIES

Aspects of Memory Capacity and Confidence in Contingency Judgements

Mélanie Clément

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements for the PhD degree in Psychology

School of Psychology
Faculty of Social Sciences
University of Ottawa

Your file Votre référence
ISBN: 0-494-10960-2
Our file Notre référence
ISBN: 0-494-10960-2

To James, with all my love.

Your love, support and encouragement in these final stages of my thesis gave me the needed strength to achieve my goal. I am looking forward to our future together.

Acknowledgements

Words cannot express my gratitude towards my thesis advisor, Dr. Pierre Mercier. I cannot imagine a better choice of person to have guided me in this process. I would like to thank him for his availability and his undying patience with my many questions. I will always remember his encouraging words and his insightful remarks, especially in these last stages of my thesis, but also throughout the course of my doctoral studies. Time and again, his help went above and beyond what is to be expected of a thesis advisor. I would also like to thank him for his faith in me and for teaching me the necessary skills to become a great researcher. This wouldn't have been possible without him.

I would also like to thank the members of my thesis committee: Dr. Alain Desrochers, Dr. Pierre Gosselin and Dr. Catherine Plowright for their helpful input and feedback in my thesis. Having all taught me at one point or another, you have contributed in some way in making me who I am today and for that, I am grateful. I would also like to thank my external evaluator, Dr. Claudette Fortin, for her insightful comments and questions.

Finally, I would like to thank my family and friends for their support, patience and understanding while I was busy with my studies. I will make it up to you!

Abstract

Theories of contingency judgements generally agree that 1- memory is a structure that possesses a limited capacity and that 2- it plays an important role in the detection and assessment of covariations. Empirical evidence, although limited in the specific context of contingency judgements, seems to support these notions. While theorists agree that some information needs to be held in memory in order to reach a contingency judgement, they disagree, however, on the exact type of information. As a result, they offer different predictions as to what would increase memory load as involved in contingency judgements. Kareev (1995, 1997) implicitly assumes that people attempt to memorize the sequence of events leading to a contingency judgement and, therefore, the longer the series of events, the higher the memory load. On the other hand, Wagner (1976, 1981; Rescorla & Wagner, 1972) proposes that people base their judgement on the strength of a "mental bond" rather than the recall of the series of episodes. In this view, the manipulation that taxes memory capacity is not the length of a series of events but the presence of multiple simultaneous contingencies. The current thesis aimed at clarifying the role that memory plays in the assessment of covariations by contrasting these two opposing viewpoints. Five experiments examined the role of memory capacity in contingency judgements by means of: 1- increasing the length of the series of single events experienced; 2- increasing the number of contingencies presented simultaneously; 3- examining the effect of individual memory capacity. Results generally support Wagner's theory with additional findings falling outside of the theory's explanatory power.

Table of contents

Introduction

Relations between events in our environment are numerous and can affect people's course of action in a number of situations. For example, the relation between grey clouds and rain leads us to take an umbrella along on a cloudy day or, more importantly, the relation between certain atmospheric conditions and a hurricane helps inhabitants of an affected area to prepare for the worst. This ability to detect predictive relations between stimuli guides beings into seeking out events that are reliable signals of rewards (e.g., exercising in order to have good health) and avoid those that signal threats (e.g., quitting smoking to avoid lung cancer). In addition to guiding people's actions, relations between events or *event covariations* are also helpful when making categorical judgements (Shanks, Lopez, Darby & Dickinson, 1996). For example, after observing a certain constellation of symptoms in a patient, a doctor will make a diagnosis based on which disease those symptoms have been known to covary with. Furthermore, researchers have also raised the importance of the detection of the relation between an antecedent and a consequent as crucial for the process of scientific reasoning (Arkes and Harkness, 1983), as well as in the social attribution of causes to our own and others' behaviours (Kelley, 1967, 1971). In short, perception and assessment of relations in our environment are essential to humans' adaptive functioning as they provide "a means for explaining the past, controlling the present, and predicting the future" (Alloy & Tabachnik, 1984, p.112). The capacity to create associations between different stimuli may well lie at the basis of learning.

Given its importance in people's lives, it should come as no surprise that evidence for the detection of covariations has been observed as early as the child's first week of

life. DeCasper and Fifer (1980) have shown that, three days after birth, babies can learn to modify their sucking rhythm in order to hear the sound of their mother's voice. Two days old infants will also learn to control their sucking in order to hear speeches in their native tongue (Moon, Cooper, & Fifer, 1993). Similarly, infants aged one to seven days old can rapidly learn to alter their oculomotor response in order to hear music (Darcheville, Madelain, Buquet, Charlier, & Miossec, 1999). Human beings are not only excellent at detecting contingencies at any age, but most adults are also quite accurate at intuitively assessing the strength of such relations. Sensitivity to variations in contingencies has been observed in studies involving a single action and its outcome (Alloy & Abramson, 1979), in cue-outcome learning tasks (Shanks, 1991), in experiments using a response rate as dependent measure (Chatlosh, Neunaber, & Wasserman, 1985), and between continuous variables (Beach & Scopp, 1966). While many studies have focused on human's ability to detect and assess contingencies, as well as the ways in which they might proceed to accomplish such a task, few have explored the role of memory and memory capacity as involved in contingency judgements. The current thesis is aimed at investigating some of these issues.

The Study of Contingency Judgements

Covariations are ubiquitous, and much research (see Allan, 1993; Alloy & Tabachnik, 1984; Crocker, 1981; Shaklee, 1983 for reviews) has been devoted to understand how the human cognitive system is designed to perceive them. The paradigms used to study covariation perception vary greatly. The variables entering into a contingency are often an action, real or imagined, and its outcome (e.g., using a certain ingredient to make a cake rise or not, Shaklee & Mims, 1986; firing shells in an attempt

to explode a tank in a video game, Shanks, 1985), or relating an event and a stimulus (e.g., sunny weather / skin smoothness, Arkes & Harkness, 1983; symptoms and diseases, Shanks, 1991). Although the relation can be causal, it need not be so. In many cases, the events are binary and they can be cast in a 2 x 2 contingency table where cell *a* contains the frequency of occurrence of the presumed cause with the outcome, cell *b* contains the frequency of the presumed cause without the outcome, cell *c* contains the frequency of the outcome without the presumed cause, and cell *d* contains the frequency of the joint absence of the presumed cause and the outcome. The 2 x 2 contingency table is represented in Figure 1.

Insert Figure 1 approximately here

The relation between two such binary events is commonly quantified by researchers using the ΔP coefficient (Allan, 1980), which is the difference between the probability of the outcome given the presumed cause *[P(O|C)]* and the probability of the outcome in the absence of the presumed cause *[P(O|~C)]*. This is formally represented in Equation (1), which relates the probabilities to relative frequencies.

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

Although researchers often use the ΔP coefficient to assess the statistical value of a contingency, they clearly disagree on whether the cognitive assessment of binary covariation is based on this coefficient or not. While it is still unclear how people arrive

at a contingency judgement, their assessment is nonetheless often surprisingly close to the result of the ΔP calculation.

However, even though human beings generally excel at estimating contingencies, there exist a number of cases where their judgements differ from the norm (i.e., deviate from ΔP). For example, biases have been observed in judging null contingencies in cases where the outcome occurred frequently (e.g., Baker, Berbier, & Vallée-Tourangeau, 1989; Chatlosh, Neunaber, & Wasserman, 1985; Dickinson, Shanks, & Evenden, 1984) or represented an incentive such as a monetary gain (e.g., Alloy & Abramson, 1979). The importance of this basic process of detecting or establishing connections between different events, actions or stimuli, is further emphasized in studies that show that the presence of biases in contingency judgements are sometimes associated with negative consequences. For example, Seligman (1975) attributed depression to an inability to perceive the relationship between one's behaviour and its subsequent beneficial consequences. Similarly, Hamilton (1976) stated that stereotypes may arise towards minority groups when the relationship between their group and negative behaviours is overestimated. A successful model of how the mind works need not only be accurate in cases where people's judgements are accurate but also needs to predict when and why people's judgements sometimes fail.

Contingency judgements models can be broadly grouped into two categories: statistical and associative. Statistical models presume that people accumulate and transform the frequencies of the four basic events in the 2 x 2 contingency table, on which they calculate ΔP or its equivalent. One such model is the probabilistic contrast of Cheng and Novick (1990, 1992) later expanded into the causal power theory, also known

as the power PC theory (Cheng, Park, Yarlas, & Holyoak, 1996; Cheng, 1997). Although this model is complex, its core computation is that of ΔP at the numerator. Another model in this class is Busemeyer's (1991; Busemeyer & Myung, 1992) information integration theory adapted by Kao and Wasserman (1993). This system, formally described in Equation (2), weighs the frequencies of each cell of the contingency table.

$$R_w = \frac{W_A a - W_B b - W_C c + W_D d}{W_A a + W_B b + W_C c + W_D d} + error \qquad (2)$$

Associative models are strikingly different from statistical ones. Their main difference is that they do not require the accumulation and transformation of cell frequencies. They include such instances as Rescorla and Wagner's (1972) competitive cues, Wagner's automatic memory processing, otherwise known as the standard operating procedures (SOP) theory of memory (1976, 1981; Mazur & Wagner, 1982; and its extension into AESOP, Wagner & Brandon, 1989), and Gluck and Bower's (1988) adaptive network. Their main assumptions are that: 1) every exposure to one of the four basic events related to the presumed cause-outcome pair will result in either an increment or a decrement in the strength of a "mental bond" between the elements of the internal representations of the stimuli in the pair; 2) the change in associative strength that occurs on each trial is a function of the difference between the existing strength and the maximum possible for that pair.

Among the associative class, the model that has been the most discussed, empirically tested, and most frequently contrasted with other classes of models is that enounced by Rescorla and Wagner (1972), also known as the Rescorla-Wagner model, along with its expansion into the SOP theory (Wagner, 1981). Although originally designed to account for learning in animal conditioning tasks, the Rescorla-Wagner

model was later extended to contingency judgements in humans (e.g., Shanks, 1985; Miller, Barnet, & Grahame, 1995; Siegel & Allan, 1996). This model is now known for its widespread influence and its application in various fields of study (e.g., verbal learning, category learning, reasoning, social psychology) (Siegel & Allan, 1996). Its later expansion into SOP made it the only theory to detail the memory processes as involved in the computation of covariations. Due to its detailing of the memory processes involved, SOP is the prime candidate for testing predictions involving memory effects. Because the aim of the current thesis is to investigate the memory aspects as they relate to contingency judgements, the Rescorla-Wagner model and its memory extension (under SOP) will therefore be retained as the representatives for the associative class.

Similarly, since Cheng and Novick's probabilistic contrast model and its evolution into the causal power theory have also generated vast interest and motivated a number of studies, they will be retained as the representatives for the statistical class. However, since they do not embed a memory mechanism, they will not be tested per se. Instead, it will be presumed that their memory view would be compatible with the work of Kareev (1995, 1997), as will be presented later. The following sections present the theories in greater details.

The Rescorla-Wagner Model and the SOP Theory

In their original model, Rescorla and Wagner (1972; see also Wagner & Rescorla, 1972) postulated that events become associated with one another because of an *associative strength* that forms between the internal representations of these stimuli when they are paired. The associative strength between stimuli will vary depending on experience: It will strengthen when events are presented together and weaken when the

events are presented separately. The strength of this associative link will gradually form through repeated presentations of the stimuli and eventually reach an asymptote, which represents the maximum associative strength possible between the two stimuli. Rescorla and Wagner proposed that there is a fixed or limited amount of associative strength that exists for any given outcome event. Therefore, in a case where there is more than one stimulus paired with an outcome, all stimuli present on that trial will "share" the available associative strength: "the effect of a reinforcement or nonreinforcement in changing the associative strength of a stimulus depends upon the existing associative strength, not only of that stimulus, but also of other stimuli concurrently present" (Rescorla & Wagner, 1972, p. 73). This phenomenon has been referred to as "cue competition". Cue competition occurs with compound stimuli (e.g., a light and a sound both announcing the arrival of food), but also in most cases where only one stimulus is presented, as the model also accounts for associations formed with the context in which the stimulus is presented. In addition to cue competition, changes in associative strength on a given trial will also depend on the associative strength already accumulated during previous trials.

These assumptions are embodied in Equations (3) to (6), from Rescorla and Wagner, where α and β are stimulus salience parameters for cues (A or X) and outcomes (present, o, or absent, $\bar{o}$), λ is the maximum strength possible ($\lambda_O = 1$, $\lambda_{\bar{O}} = 0$), ΔV is the change on a given trial, and V is the cumulative strength. A represents the target presumed cause, X the background of all known and unknown alternate candidate causes (i.e., the context), and V_{Sum} is the total associative strength of all cues present at a given time.

$$\Delta V_{An} = \alpha_A \beta_{O or \bar{O}} (\lambda_{O or \bar{O}} - V_{Sum n-1}) \tag{3}$$

$$V_{An} = V_{An-1} + \Delta V_{An} \quad (4)$$

$$\Delta V_{Xn} = \alpha_X \beta_{O or \bar{O}} (\lambda_{O or \bar{O}} - V_{Sumn-1}) \quad (5)$$

$$V_{Xn} = V_{Xn-1} + \Delta V_{Xn} \quad (6)$$

The presence of stimulus salience parameters for the cues in these equations (i.e., the alphas) indicates an acknowledgement that some stimuli may acquire associative strength at different rates despite equal reinforcement. The authors propose that the associative strength of a compound, V_{AX}, can be calculated by adding the associative strength of each component present on the given trial: $V_{AX} = V_A + V_X$. Rescorla and Wagner also allow both for excitatory associations (as indicated by a positive V resulting from the equation) and inhibitory associations (as indicated by a negative V) to be formed between stimuli.

Wagner later expanded on these assumptions in proposing the standard operating procedures or SOP[1] theory of memory (1976, 1981; Wagner & Brandon, 1989) in which he detailed the memory processes involved in the formation of associative strength and therefore, as involved in contingency judgements. According to SOP, any given stimulus is represented in people's memory in the form of a *node*. In the spirit of stimulus sampling theory (Estes, 1955a, 1955b), a node is defined as a collection of elements making up the representation of a stimulus (e.g., information regarding the stimulus's size, colour, texture, shape, etc.). Given the different terminology used by theorists to define the processes or elements of memory (e.g., working memory, short-term memory)

[1] In addition to being an acronym for "standard operating procedures", because of its similarities with the opponent-process formulation of Solomon and Corbit (1974) and Schull (1979), Wagner (1981) stated that the acronym could also be used to represent a "sometimes opponent-process" theory of memory.

and the controversy surrounding such distinctions, Wagner opted for a more neutral distinction between a state of inactivity (*I*) and two states of activity – a primary state (*A1*) and a secondary state (*A2*). According to Wagner, memory nodes can be in any of these states at a given time.

As originally postulated by Atkinson and Shiffrin (1968), Wagner also assumes that the memory system interacts with the environment through a *sensory register*. In general, a memory node will be in its state of inactivity until the presentation of its corresponding stimulus to the sensory system. This prompts the node into its primary state of activity (*A1* state) in which the stimulus will be *rehearsed* (Atkinson & Shiffrin, 1968). The activated representation will then gradually decay to the secondary state of activation (*A2* state) to finally return to its state of inactivity (*I*), according to different probabilities of decay (*p*). Because of memory limitations (as will be detailed shortly) the probability of decay between the *A1* and *A2* states is assumed to be greater than that between the *A2* state and the inactivity state. Once inactive, a node will remain in that state until its corresponding stimulus is once again presented to the sensory system.

Wagner supposes that when the representations of two stimuli are *simultaneously* in their primary state of activation (*A1*), an excitatory link is formed between these two nodes and associative strength forms. Elements that are already in their state of activation (e.g., through priming) when presented to the sensory system a second time will not be activated to the degree that they would otherwise be. Nodes can also be activated through the spread of activation from another activated node that has previously been associated with it. However, nodes that are activated in this fashion will be brought directly into *A2*

state, rather than in their *A1* state. Inhibition will be obtained between two stimuli when one is in its *A1* state while the other is in its *A2* state.

In accordance with the idea that memory capacity is limited, Wagner also proposes limitations on the *number of nodes* that can be simultaneously active. Indeed, Wagner postulates that the number of nodes that can be simultaneously in their primary state of activation (*A1*) is very limited, such as possibly only two or three nodes could be *fully* in this state simultaneously. As each memory node is composed of elements, Wagner uses the constant *C1* to represent the summed proportion of elements that can be simultaneously in their *A1* state across all memorial nodes. When more than two or three nodes are in their *A1* states simultaneously, in order to keep the proportion of elements in their *A1* states equal to *C1*, as the number of activated nodes increases, this momentary increase must be "compensated for by some decrease in the proportion of *A1* elements over all nodes so that the summed proportion will still equal *C1*" (Wagner, 1981, p. 23). In order to accomplish this, the probability of decay applied to a proportion of *A1* elements in each node in the system will be momentarily increased (i.e., part of the elements contained in each node will decay faster in order to keep the summed proportion of all active elements to *C1*).

Wagner proposes less severe limitations on the *A2* state. The capacity to maintain elements in their secondary state is, according to him, higher such as 10 or 15 nodes could be fully in this state simultaneously. However, if this capacity is exceeded, the same mechanism will be applied as for the elements in *A1* state: In order to keep the proportions of elements in *A2* state equal to a *C2* constant the probability of decay (to the inactive state) for a portion of elements across all nodes will be increased. While Wagner

does not attribute values to the two constants (*C1* and *C2*), he proposes that *C2* should be approximately 5 times the value of *C1*. Wagner formally represented these concepts through the distractor rule. The distractor rules for the *A1* and *A2* states are represented in Equation 7 and 8, respectively.

$$p'_{d1} = p_{d1} + \frac{\Delta p_{A1,X}}{C1} \tag{7}$$

$$p'_{d2} = p_{d2} + \frac{\Delta p_{A2,X}}{C2} \tag{8}$$

In these equations, p'_d represents the higher momentary probability of decay, p_d represents the decay probability, $\Delta p_{A,X}$ is the momentary increase and *C*, the memory capacity constant.

One phenomenon that has been seen to pose a problem for standard elemental theories (i.e., in which memory nodes are seen as a collection of elements), such as the Rescorla-Wagner model (1972), is their inability to account for contextual sensitivity. While the model includes a calculation of the strength between the context and an outcome in the total associative strength of all cues presented on a given trial (i.e., through cue competition), it does not predict that the context will have an influence on the formation of associative strength of other stimuli presented on that trial (when the memory limitations are not exceeded). In brief, the Rescorla-Wagner model, as originally formulated, predicts that "the associative strength [...] is unaffected by the context in which it is presented" (Pearce, 1987, p. 65). However, we now know that this is not the case. Many studies show evidence that the context does have an impact on learning. For example, studies of human memory have shown that recall of material learned in one context can be impaired when tested in another (e.g., Baddeley, 1976; McGeoch, 1942).

In order to circumvent this problem, Wagner (2003) introduced the *replaced elements conception* as an extension to his original model. The replaced elements conception addresses the effects of context on other stimuli by introducing the presence of *context dependent elements* within each node. These will be activated or not depending on the presence or absence of other concurrent stimuli (i.e., making up the context). According to this view, the "representation of a compound of stimuli involves the replacement of some elements otherwise contributed by the constituent stimuli in isolation" (Wagner, 2003, p. 13). In other words, when a stimulus is presented within a given context, some of the elements contained in the stimulus's node are replaced by some of the elements from the context's node. Because of this exchange of elements, this would account for the influence of context on associative strength of a given contingency (since some of the elements from the context node are also included in its formation). This conception, although not directly linked to memory capacity as studied in the current thesis, will be useful to clarify results obtained with studies using multiple contingencies. It will therefore be revisited in greater details later.

The Probabilistic Contrast Model and the Causal Power Theory

According to the probabilistic contrast model (Cheng & Novick, 1990, 1992; Cheng & Holyoak, 1995), the process of detecting associations between stimuli does not happen through increasing and decreasing mental associations. Instead, their model is based on an analogue of statistical contrasts that require the estimation and comparison of two proportions. They propose that people arrive at their judgements by evaluating the difference between the conditional probability of the outcome given the presence versus the absence of the target predictor (i.e., the ΔP rule). Indeed, studies have shown that

people are able to assess those two probabilities with reasonable accuracy (Alloy & Abramson, 1979; Robinson, 1964; Shuford, 1961).

Another main difference between their model and that of Rescorla-Wagner is that it is also a causal model. While the Rescorla-Wagner model focuses only on contingency assessment, the probabilistic contrast model also seeks to understand how humans make causal attributions. According to the probabilistic contrast model, causes can be facilitatory (positive difference between the two proportions) or inhibitory (negative difference between the two proportions). The authors assume that, in order to be perceived as causal, a stimulus must be perceived as preceding or co-occurring with its effect. A stimulus that has been identified as a potential cause is then evaluated by its contrast calculated over a *focal set*. A focal set is the set of events that consist of the context in which a potential cause is embedded. According to the authors, the focal set used will often not consist of the universal set of events. When assessing a contingency, people will select which cues will be included in a focal set, based on their prior experience and circumstances. Using frequencies of events (as represented in a 2 x 2 contingency table), people estimate the probabilities with which an effect e will occur in the presence and in the absence of its potential cause i. According to the authors, if ΔP is noticeably different from 0, it is perceived as a cause. In other words, a cue is only viewed as causal if its presence makes a difference to the probability of the effect.

The probabilistic contrast model also allows for multiple causes to co-exist. Any stimulus i whose presence increases the probability of obtaining an effect e is a potential cause of the effect. If many stimuli are observed to increase the probability of the effect, then each of them are a cause of the effect. Stimuli can also combine in a non-

independent way to produce the effect. The model also differentiates between a cause and an enabling condition: An event will be an enabling condition for a cause if it is not followed by the effect when the cause is absent, but the cause is only followed by the effect in circumstances in which this event is present.

The probabilistic contrast model was later expanded into the causal power theory (Cheng, 1997; Cheng, Park, Yarlas & Holyoak, 1996). Following the observation that not all covariations with a ΔP different from 0 necessarily indicate causal relationships (for example, sunrise may follow a rooster's crow but the rooster's crowing does not cause the sun to rise), Cheng (1997) posited that causes are inferred not only when a certain degree of covariation is detected but also when people can detect the presence of *causal power* between the stimuli. Cheng defines causal power as "the intuitive notion that one thing causes another by virtue of the power or energy that it exerts over the other" (p. 368). According to this view, causes are not only followed by their effect but actually produce or generate them. Hence, the crowing will not be seen as the cause of sunrise because there is no mechanism through which sound causes objects to elevate. Although this model is complex, its core computation is that of ΔP at the numerator as shown in Equations 9 (for generative causal power) and 10 (for preventive causal power).

$$p_i = \frac{\Delta P_i}{1 - P(e \mid \bar{i})} \tag{9}$$

$$p_i = \frac{-\Delta P_i}{P(e \mid \bar{i})} \tag{10}$$

In these equations, p_i represents the power of a cause i to cause an effect e, ΔP_i is the contingency between candidate cause i and effect e and $P(e|\bar{i})$ is the probability of e given the absence of i. The causal power theory also accounts for why people prefer to

use different focal sets when calculating their proportions. That is, people prefer to compute a contrast conditional on the absence of all other plausible causes to assess whether a candidate cause is generative ($P(e \mid \bar{i}) \cong 0$) and they prefer to compute a contrast conditional on some generative cause being constantly present to assess whether a candidate is preventive ($P(e \mid \bar{i}) \cong 1$). We can see from Equations 9 and 10 that when these conditions are met, ΔP will become a close estimate of causal power, in each of these cases. This helps people get a better estimate of causal power and hence, can account for their preferences.

Cheng and Novick (1990) also acknowledge the reality of memory limitations and state that people will have greater difficulty processing more complex interaction contrasts due to those limitations. They propose that if the complexity is great enough, it may even become impossible for people to do so. However, they do not detail these memory limitations. In talking about memory capacity limitations, the authors say: "Because the probabilistic contrast model is a computational model [in Marr's, 1982, sense of the term] that specifies *what* is computed, rather than a process model that specifies *how* the computation is carried out, it does not deal with such processing limitations" (Cheng & Novick, 1990, p. 550). Since the authors do not define specifically what would contribute to increase the complexity of contrasts, it is not sure what the model's predictions are regarding the best way to test memory capacity. Because the authors do not detail the memory processes involved either, it is difficult to explain how they would account for a greater processing difficulty or what effects should be expected from these difficulties. Because of this, Cheng's model, as previously stated, does not

really allow for a discussion of the impact of memory limitations on covariation judgements.

Memory and Contingency Judgements

It is now a commonly accepted concept in psychology that memory is a structure that possesses a *limited capacity* and a number of studies corroborate this idea (e.g., Baddeley, Thomson, & Buchanan, 1975; Miller, 1956). In accordance with this proposition and if memory does play an important role in the assessment of covariations (e.g., Mercier & Parr, 1996; Parr & Mercier, 1998; Shaklee & Goldston, 1989), it only follows that people's judgements should be affected in situations where memory load is increased. To this day, few studies have investigated the role of memory in contingency judgements. However, in the studies that have, preliminary evidence for the role of memory can be found. Indeed, studies that tax memory capacity have demonstrated impairment in the assessment of covariations.

For example, empirical studies have revealed judgement impairment in trial-by-trial or "on-line" methods of stimulus presentation, 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). It was suggested that judgements decrease in accuracy because trial-by-trial presentations of stimuli might lead to a heavier memory load than "off-line" methods of data presentation, where the contingencies are presented in the form of tabled frequencies or numbers in a story. Although interesting, this explanation remains suggestive, as off-line and on-line experiments differ in more ways than one. For instance, it is possible that the off-line method may encourage participants to perform a calculation similar to ΔP, also

explaining why their judgements may be closer to the normative answer. The increase in memory load was also invoked in support of experiments showing lower accuracy in judging contingencies with continuous, as opposed to binary, stimuli (Trolier & Hamilton, 1986). It is argued that, because continuous stimuli are more complex, they are also more demanding to process. However, once again, this argument remains suggestive.

A stronger account of the impact of an increased memory load was obtained in a study by Mercier and Parr (1996). In this study, participants were asked to assess the degree to which paint would protect or harm a tank against "visually guided" alien mines. Processing difficulty was manipulated through stimulus duration (50 or 200 ms) and inter-trial intervals (ITIs) (50, 200 or 1000 ms). Shorter stimulus duration and inter-trial intervals increase the difficulty of the task by forcing the cognitive system to acquire and maintain information faster, hence increasing the load on memory. Judgement accuracy declined towards zero with shorter durations and shorter ITIs. Parr & Mercier (1998) also replicated this effect with older adults.

While general models of memory all predict a decrease in accuracy when memory capacity is surpassed (and this is what is observed), we need to turn to contingency judgements models for a more detailed account of the effect of an increased memory load on contingency judgements per se. Unfortunately, as previously stated, the probabilistic contrast model and its expansion into the causal power theory do not make clear predictions regarding processing limitations, as they aim to specify *what* is computed, rather than *how* the computation is carried out (Cheng & Novick, 1990, 1992; Cheng, 1997). The SOP theory, on the other hand, does offer such predictions. As seen previously, Wagner predicts that when memory capacity is surpassed, elements in their

active states will decay faster, across all memorial nodes. Since a proportion of the elements in their active state, over all nodes, will decay faster (to keep the proportion of all active elements equal to a constant), the links formed between the nodes should be correspondingly weaker, as this quick decay will prevent the normal growth of the associative strength (it will not be able to reach asymptote as quickly as it should). Relations between stimuli should therefore be *underestimated* in those cases (comparatively to situations in which memory capacity is not exceeded). This is indeed what was observed in previous studies (e.g., Mercier & Parr, 1996).

While these results offer preliminary insights on memory capacity as involved in contingency judgements, the issue is still far from being clearly understood. To further complicate matters, researchers studying contingency judgements do not agree on the best way to tax memory capacity. Theorists of contingency judgements agree that, in order for covariations to be perceptible, some information must be memorized. They disagree, however, on the type of information that needs to be remembered and hence, disagree on the most appropriate way to increase memory load. Explicitly or implicitly, some researchers believe that the specific, sequential, events involved in a contingency must be remembered in order to assess it. Others view the memory of specific events as much less necessary as the memory of some form of cumulative information.

Kareev's Narrow Window Approach

The issue of memory capacity and its role in covariation detection came to the fore in two theoretical papers by Kareev (1995, 1997). In his "narrow window" approach, Kareev implicitly assumes that people memorize the individual episodes of events leading to the detection of a covariation. Since memory capacity is limited, the length of

the series that can be memorized is also limited. Consequently, Kareev argues that people determine the existence of a relation between two variables by relying on *samples* of events where the variables are either present or absent together, or where one variable is present and the other is absent. In addition, because the capacity of working memory varies from one individual to another, the size of the sample considered should also vary across individuals. As the sampling distribution of the Pearson correlation coefficient tends to be skewed when the correlation of the population is different from zero (Hays, 1963) and the skewness increases with reductions in sample size, Kareev argues that the sampling of a reduced number of events, or smaller working memory capacity, should lead people to *overestimate* the normative relations between events. Given these considerations, it seems that Kareev assumes that 1- the information contained in the individual episodes enters the cognitive system where 2- it is integrated by computing Pearson's r or an equivalent such as ϕ, φ, or ΔP^2.

In contrast with Kareev's view, other models of contingency judgements postulate that it is not necessary to remember the individual episodes of events that constitute a covariation. As seen previously, according to statistical models such as Cheng's (1996), in order to assess a covariation, people would only need to remember the four frequencies involved in the 2 x 2 contingency table (to use in the calculation of the two probabilities). Therefore, for a single cause-outcome pair, the minimum memory resources required by statistical models are four basic cumulative frequencies in memory as they enter in the

---

[2] Pearson's r, ϕ, φ and ΔP are isomorphic under the restriction that the shape of the distribution be similar for the two binary variables.

computation of the covariation index. These memory resource requirements remain constant regardless of the size of the sample of events considered. It is important, however, to note that the probabilistic contrast model or the causal power theory do not propose that the individual episodes will necessarily be completely forgotten either, but only that they are not required for evaluating a contingency.

While Kareev's narrow window approach predicts an overestimation with smaller samples of events, because the probabilistic contrast model or the causal power theory do not detail the memory processes involved in contingency judgement, it is difficult to determine what they would predict in this particular case. According to statistical models, it can be derived that the average estimate should not change systematically with sample size when the contingency remains constant. The estimates might fluctuate more, in both directions, reflecting the larger variance of the smaller samples, but there is no provision in the model's core formulas to handle this. This prediction created a problem for the probabilistic contrast model, as it was unable to account for the well-known phenomenon of the learning curve. In order to circumvent the model's inability to predict the traditional learning curve, Cheng and Holyoak (1995) added to their model the assumption that "confidence in the assessment of a contrast is presumed to increase monotonically with the number of cases observed" (p. 273). They argued that people's judgements are indeed constant with trials, but that their confidence in these judgements increases across trials. While the authors did not elaborate on the concept of confidence, according to them, acquisition curves are observed as a result of combined contingency and confidence estimates.

While this added element allows the model to account for lower estimates in smaller samples (as reflected in the learning curve), a second prediction can also be derived from it. Indeed, even though statistical models differ from Kareev's approach on the type of information that needs to be remembered in order to arrive at a contingency judgement, they could also lead to the same prediction (i.e., overestimating with smaller samples) as the distributions of the coefficients they propose (e.g., ΔP) are also skewed for small samples. Apart from the requirement of episodic memory, the computational nature of Kareev's model makes it a member of the statistical category. If Kareev's predictions are accurate, his suggestion would be entirely compatible with Cheng's model (as they both rely on a process that calculates, at its basis, a statistical coefficient).

Although associative models are sensitive to the frequencies of events, they do not require that they be tracked cumulatively as statistical models do. They do not require the memorization of the specific events (although they do not preclude it) as proposed by Kareev, either. As mentioned previously, associative models only require the recall of the previously accumulated strengths (V_A and V_X), and their combinations with the change ($\Delta V_{A or X}$) on the current trial. Thus, ignoring additional requirements for weights and salience parameters, the associative models can operate minimally with a three-register memory capacity. Contrary to Kareev's narrow window approach, Wagner's (1981) model predicts that contingency judgements should *decrease*, rather than increase, with smaller sample sizes. This prediction is not based on the limitation in working memory capacity to store the string of episodes relating the contingent events, but rather on the fact that, with a small number of episodes, associative strength cannot yet reach asymptote (λ).

Which prediction is correct concerning the effect of stimulus sample size on contingency judgement is a matter of empirical testing. Unfortunately, the existing empirical reports are contradictory. Kareev, Lieberman & Lev (1997) reported two experiments that, they claimed, supported Kareev's (1995) narrow window approach. In the first experiment, they tested memory capacity but not sample size. Participants were divided into groups corresponding to two levels of working memory capacity as determined by the median score on a standard digit-span test. Participants were asked to predict the content (printed symbols X or O) of 128 envelopes (coloured red or green). After each prediction, participants evaluated their predictions by verifying the envelope's content. The correlation between the colour of the envelope and its content varied between -.60 and +.60 across conditions. The participants' perceived correlations (φ) were calculated from their predictions on each trial. The authors reported that the perception of correlation was "more extreme" among participants with low working memory capacity as compared with participants having high working memory capacity. Note that this interpretation relies on a comparison between the low and high memory capacities rather than a comparison against the normative value. This is an issue that will be revisited later.

However, independently of the basis for comparison, the interpretation of Kareev et al.'s (1997) data remains uncertain, as the participants' perceptions of correlation were not collected. A contingency coefficient φ was calculated based on the participants' predictions of each envelope's content. This is equivalent to computing contingency from the participants' estimates of the probability of an X given that the envelope was red *[P(X|Red)]* or green *[P(X|Green)]*. One problem with this approach is that Wasserman,

Elek, Chatlosh, and Baker (1993) have already shown that there is not a good match between probability estimates and ΔP (ΔP calculated from the participants' probability estimates as compared to the real normative ΔP), which is equivalent to φ, nor between probability estimates and direct contingency estimates (i.e., asking people to evaluate the contingency). An additional problem is that since we do not know for sure how the information is integrated computationally in the cognitive system, any test of people's judgements (the trial predictions) embedded into statistical calculations performed by the experimenter (the φ coefficient) is open to detecting biases that may not be part of the cognitive operations since these operations may not include φ or its equivalent. In other words, the bias observed here may be part of the calculated correlations but it is quite unclear if it is part of the subjective perception at all.

Furthermore, since the participants were merely asked to make single trial predictions, it is hard to see how memory capacity could play any role in this experiment. Even if participants remembered the individual envelopes sampled previously, this information would be of no help in predicting the content of the next trial. Only the two conditional probabilities matter and only two memory registers are needed per probability to keep track of it: The cumulative frequency of red, and the cumulative frequency of X given red, each of which can be updated on each trial. This is not to say that the process is error free but whatever amount of error is entailed, it is not clear how it should systematically inflate correlation estimates for small samples.

In a second experiment, Kareev and colleagues assessed the relation between both sample size and memory capacity on a single predictive judgement. Participants were presented with a sample of cards drawn from a population of 100 cards. Each card

depicted two different lines, with the correlation between the length of one line and the angle of the other line maintained constant across different sized samples. The size of the sample was either smaller by two, the same value, or larger by two than the participants' memory capacity as measured with the digit-span test. The participants made one prediction for each sample size. The authors reasoned that if the perceived correlation was high, then participants should tend to make predictions closely related to the predictor and, on average and after standardisation, there should be less absolute discrepancy between prediction and predictor ($|Z_{\bar{Y}} - Z_X|$) in the small sample condition. In addition to sample size, the authors also manipulated whether the sample stimuli were entirely visible or not at the time of the prediction. They reasoned that people might have a natural distrust of small samples but that this distrust might disappear when memory load is eased by making the cards visible. However, they did not explicitly state in what direction the judgements might change as a result of this double process involving reduced memory load and increased trust.

On the surface, the results regarding the sample size manipulation appeared consistent with the authors' expectations when the sample was invisible: The smaller the sample, the lesser the discrepancy ($|Z_{\bar{Y}} - Z_X|$) between the prediction (participants' average guess of line length) and the predictor (line angle), hence the authors' inference of a high perceived correlation between predictor and criterion (actual length and angle – length and angle were counterbalanced across participants). However, in order for these results to follow from the narrow window approach, they must coincide with an actual larger correlation in the small samples because, according to the approach, it is the objective sampling bias itself that leads to higher judgements. Unfortunately, the analysis

of the actual sample correlations showed that all *F* values in a sample size by sample presence ANOVA were "far from significant" (pp. 284-285). This indicates that whatever was influencing judgements, it was not what the narrow window approach expected.

The opposite result was found when the sample was visible: The smaller the sample, the larger the value of $|Z_{\bar{Y}} - Z_X|$, hence the author's inference of a lower perceived correlation between predictor and criterion. Whether or not this result is in line with the narrow window approach is difficult to decide because of the lack of a clear prediction by the authors. If people have gathered from previous experience that small samples are variable, then this should be a property related to size rather than visibility. Given that a small sample remains small even when it is visible, it is unclear how the visibility of the stimuli could remove distrust (as proposed by the authors). If trust is ignored, visibility should reduce memory load in Kareev's (1995) narrow window approach by removing all need to memorize the individual stimuli, leaving only the effect of sampling bias to influence judgement. If this was the case, the pattern of results should be the same for visible stimuli as for invisible stimuli, perhaps even more so, because visibility should reduce forgetting. If we grant that distrust is somehow lifted by visibility, then participants should have trusted the data and again, via the theoretical biasing effects of small samples, they should have judged the correlation higher in the small samples than in the larger ones, manifesting this by a small $|Z_{\bar{Y}} - Z_X|$. Accordingly, the observed large $|Z_{\bar{Y}} - Z_X|$ in the visible condition contradicts the theoretical expectations with or without trust. Considering all these problems, it is difficult to determine what role, working memory capacity plays in contingency judgement. Kareev

et al.'s results become even more puzzling in the context of other evidence running contrary to Kareev's predictions and in favour of Wagner's.

In a paper that inspired an entire paradigm, Shanks (1985) developed a task resembling a video game in which participants fired delayed action shells at tanks that also crossed a minefield. Thus, once they entered the field, the tanks were at risk of exploding either because of a hit by a delayed action shell or because of a mine in the field itself. Participants were asked to evaluate the risk of explosion caused by the shells relative to the mines on a scale from -100 to +100. In four experimental conditions, the conditional probabilities of explosion given a shell and given a mine were .75:.25, .75:.75, .25:.25 and .25:.75 respectively for one positive, two null, and one negative contingency. There were 40 trials per condition. Participants were required to estimate the shell-explosion contingency after every five trials. They discriminated the four contingencies well and, more centrally to the current issue, their contingency estimates grew gradually towards their asymptotic values. This is suggesting that, contrary to Kareev's expectation, contingency in small samples (early judgements) is perceived as smaller than in larger samples (late judgements). However this evidence can only be taken as suggestive for two reasons. First, as pointed out by Baker and Mercier (1989), on a trial where there was an explosion, participants could not unambiguously ascribe its cause to either a mine, a shell or both, thus making the task logically unsound and jeopardising the validity of the empirical contingency estimates obtained. Second, because they occur late, judgements of the larger samples always occur after earlier estimates of the same contingency have already been made, raising the possibility that the participants' opinion may have been committed by their earlier answers. The first

problem was corrected in later work (e.g., Baker, Mercier, Vallée-Tourangeau, Frank & Pan, 1993; Shanks, 1987), but not the second.

To obtain a clean comparison of sample sizes within subjects, the sample size conditions must be randomized. Mercier and Parr (1996) did this. They asked participants to evaluate the relation between camouflage and safety in a tank video game. Contingency judgements were gathered directly on a scale ranging from -100 to +100. For constant ΔP values of -.50 or +.50, judgements tended significantly towards zero when the 2 x 2 contingency table contained only 8 events ($\overline{X}$ =.29) rather than 32 ($\overline{X}$ = .42) or 40 ($\overline{X}$ =.43). The ΔP value was maintained constant by setting the *a*, *b*, *c*, and *d* cell frequencies to 3, 1, 1 and 3; 12, 4, 4 and 12, and 15, 5, 5, and 15 respectively for the three sample sizes in ΔP = +.50. Thus, in this experiment, the objective contingency was maintained constant and yet, decreasing stimulus sample size reduced judgements. Mercier and Parr argued that these results, along with many others, are consistent with Wagner's associative perspective. Recall that according to associative models like Wagner's, people do not assess covariations directly. Instead, people estimate the relation based on the amount of associative strength linking the two events. Based on this principle, few exposure trials are expected to generate less strength and hence lead to lower estimates of correlation.

Mercier and Parr's results are strong but their interpretation is not as definitive as can be vis-à-vis the narrow window approach. First, the small "samples" were not really samples but small populations since their ΔP value was controlled via set frequencies in each cell. Another criticism concerns Mercier and Parr's interpretation in terms of associative strength. It is possible that the reduced judgements might have resulted from a

lower confidence level when the sample was smaller. Since confidence was not assessed separately, it is possible that the underlying process may have been diminished associative strength, reduced confidence, or both. Indeed, confidence would be an issue in any comparison of sample sizes and it should be measured simultaneously in any definitive experiment.

Overview of the Experiments

In order to settle the confidence and randomized sampling issues, Experiment 1 reprises Mercier and Parr's paradigm with three modifications: 1- confidence is measured separately from contingency; 2- small samples are selected randomly from a population of 40; 3- sample sizes and all other conditions are presented in randomized order throughout. To extend the generality of the replication, the chosen stimuli are medications and their effects rather than the camouflage of a tank and its safety in a video game.

While the narrow window approach assumes that people attempt to memorize the individual episodes that form a contingency and that this is what taxes memory, Wagner has a different view. Indeed, SOP also provides for memory capacity limitations, but not in the way proposed by the narrow window approach. Because, according to SOP, people do not need to remember the individual episodes, increasing the length of a series of events will have no impact on memory limitations. For a single contingency, this will still activate three nodes (one for each stimulus included in the contingency and one for the context), regardless of the length of the series. As previously discussed, the length of the series will have an impact on the associative strength, not because of its effect on memory load, but because with smaller samples, associative strength has not yet reached

asymptote. The memory capacity limitations proposed by SOP are in the form of the number of memory nodes that can be simultaneously activated. A proper test of an increase on memory load, in SOP's view, is therefore to increase the number of simultaneous stimuli or simultaneous contingencies. This will be tested in Experiment 2 and 3, where the number of simultaneous contingencies presented to participants will be increased to three and four, respectively.

Finally, it is important to note that while Experiment 1 tests the logic of the narrow window approach, through controlling sample size, it remains that in the narrow window approach, sample size is viewed as an attribute of the participant (i.e., a dependent variable rather than an independent variable). Experiment 4 and 5 will therefore test memory capacity using a standard digit span test (to separate participants in high and low memory capacity groups). In SOP's view, participants with a low memory capacity may have stricter limits on the number of nodes that can be activated simultaneously (in comparison with a high memory capacity group). We may therefore observe a difference in contingency estimates between memory capacity groups when they are presented with multiple simultaneous contingencies. Experiment 4 will test SOP's prediction. On the other hand, Experiment 5 will test the narrow window approach's prediction by presenting both memory capacity groups with a single contingency and using a fixed series of 20 events. According to the approach, participants with a smaller memory capacity will memorize smaller samples of the series of 20 events. Consequently, because the distribution of the Pearson correlation coefficient's skewness increases with reductions in sample size and that this approach implicitly assumes that people calculate either a correlation coefficient or an equivalent in order to

evaluate a contingency, people with lower memory capacity should overestimate the contingencies in the series (comparatively to the high memory capacity group).

Experiment 1

Method

Participants. Participants were 30 volunteer students (24 females, and 6 males) from the University of Ottawa (23 undergraduate, and 7 graduate students). They were aged between 18 and 39 years ($\overline{X}$ = 23).

Apparatus. The task was administered and the data collected using an IBM compatible microcomputer, equipped with a 14 inch VGA monitor and located in an individual testing room. Programming of the study was done using Microsoft Visual Basic Professional 6.0.

Stimuli. Stimuli were presented in sequential trials. Each trial pictured a fictitious medical file of a patient suffering from a disease causing skin discoloration. This task is inspired by an experiment conducted by Anderson and Sheu (1995).The medical files were represented using 10 cm by 15 cm windows, centered on the computer screen. Each file contained either the presence or the absence of a treatment, and the presence or the absence of the skin discoloration. The treatment (present or absent) appeared on the left portion of the medical file and its effect (presence or absence of skin discoloration) was shown on the right. The presence of treatment was symbolized using red or blue oval-shaped medication pills, measuring approximately 9 mm in height and 5 mm in width. The absence of a treatment was illustrated using the same oval-shaped pills with a black "X" overlaid on the pill's image. The presence of skin discoloration was represented with a round icon depicting a sick face (tongue sticking out) and was the colour of its

corresponding treatment (either blue or red). The absence of skin discoloration was represented with a round icon depicting a white smiling face. The face icons had a diameter of about 12mm. Each medical file was visible for 1 s, with 1 s inter-trial intervals.

Procedure. On entering the laboratory, participants were informed of the experiment's procedures on the computer screen as follows:

> Welcome to the Cognitive Psychology Laboratory. We are studying decision-making processes such as those used by health professionals. Imagine that you have access to the medical files of many people sick with different diseases. These diseases have many symptoms, including one in common: a skin discoloration. Different medications treating these diseases have been used in clinical trials. The problem is that the medications themselves can also increase or decrease the risk of skin discoloration. We will ask you to evaluate different medications. For each medication, the clinical trial contains the results of certain individual files. Depending on the number of people taking each medication, the clinical trial could contain the results of many or few individual files. Warning: more people taking a given medication does not necessarily mean that the medication is more or less effective. Each file may report the absence of medication, or the presence of medication, the presence of skin discoloration, or the absence of skin discoloration, in different possible combinations. In each case, the medication and the skin discoloration could be either red or blue. You will have an opportunity to warm up your diagnostic skills with six practice trials. The first three practice trials will contain many files, while the three last ones will contain less.

Participants were shown the different stimuli in all possible combinations (presence or absence of the treatment and presence or absence of the skin discoloration). These instructions were followed by six series of practice trials. The first three practices contained 40 medical files. Practice one had a contingency ΔP of .67. Using a 2 x 2 contingency table where cell *a* represents the frequency of the presence of both events, cell *b* represents the frequency of the presence of the treatment but the absence of skin discoloration, cell *c* represents the frequency of the absence of treatment but the presence of skin discoloration, and cell *d* the frequency of the absence of both events, the frequencies for the first practice series were 20, 10, 0, and 10 for cells *a*, *b*, *c* and *d* respectively. The trial order was randomized. Practice two had a ΔP of -.67 with frequencies of 0, 10, 20, 10, and practice three was a null contingency with frequencies of 10, 10, 10, and 10. Practices four, five and six contained only 4 medical files each with ΔPs of .67 (2, 1, 0, 1), -.67 (0, 1, 2, 1), and 0 (1, 1, 1 and 1) respectively. The contingencies of the series of practice trials were fixed to ensure a constant ΔP. After each practice, participants were asked to judge the efficacy of the treatment relative to the absence of the treatment based on all files presented for the clinical trial, using a scale ranging from -100 to 100, where -100 meant "extreme reduction of the risk of skin discoloration", 100 meant "extreme increase of the risk of skin discoloration", and 0 meant "absence of change in the risk of skin discoloration". Then, on the same screen, participants were asked to assess the degree of confidence they had in their answer using a scale ranging from 0 to 100, where 0 meant "no confidence", 50 meant "moderate confidence" and 100 meant "extreme confidence". Data from the practice trials were not included in the analyses. The experimenter read the directions to the participants and was

present for the duration of the practice trials to answer questions. Careful attention was given to the way instructions were presented, and questions answered, so as not to suggest any one particular means of arriving at judgements. The participants were told that they might have the impression of sometimes guessing the answers but that this impression was normal. They were also forewarned that the experimental series of trials could contain not only short and long series as seen in the practices but also some of intermediate length. Because series lengths were presented randomly, participants were asked to pay close attention to each medical file in order not to miss any information. It was also emphasised that the relations between treatments and effects were constant within a clinical trial but varied between clinical trials.

Following the practices, 32 test series of clinical trials were presented randomly to the participants. These 32 series were composed of all combinations of sample size (4, 6, 8, or 40), contingencies (ΔP of .20, .40, .60, or .80) and pill colour (red or blue). Sample sizes six and eight, and contingencies .60 and .80, fall in the ranges of sample sizes (7±2) and contingencies (> .50) demonstrated by Kareev (2000) to be crucial for the statistical overestimation of binary correlations based on φ or ΔP. The frequencies for cells *a*, *b*, *c*, and *d* used for each contingency were as follows: frequencies of 12, 8, 8, 12 for a ΔP of .20, frequencies of 14, 6, 6, 14 for a ΔP of .40, frequencies of 16, 4, 4, 16 for a ΔP of .60, and frequencies of 18, 2, 2, 18 for a ΔP of .80. These specific frequencies were chosen because they make up contingency tables for which the marginal totals were always equal. This ensures a symmetrical distribution of the presence/absence of the treatment medication as well as a symmetrical distribution of the presence/absence of the outcome, making ΔP, ϕ, and Pearson's *r* isomorphic for these tables. Samples of medical files were

then picked randomly from the total of 40 medical files for each series of trials and shown to each participant. After each series of medical files, the judgement screen appeared with the rating scales. The contingency judgement scale was reset to 0, and the confidence judgement scale to 50 (i.e., middle of each scale), for each new judgement screen. Participants could change their answers before starting the next clinical trial but they could not return to their current judgement screen after submitting their responses. Before each series, a warning window appeared reminding the participants that they were about to start a new clinical trial and to start their evaluation afresh.

Results and Discussion

All data were analysed using repeated-measures analyses of variance (ANOVA) with a Type I error rate of .05. To protect against possible violations of sphericity assumptions, a Huynh–Feldt correction was applied to all repeated measures with >1 df in the numerator. Tukey's honestly significant difference (HSD) test was used for post-hoc pairwise comparisons.

Figure 2 depicts the main results for contingency estimation. Before proceeding with the examination of the sample size and confidence effects, good judgemental discrimination must be ascertained across the experimentally manipulated contingencies (ΔP = .20, .40, .60, or .80). The ranking of the four curves in Figure 1 indicates that this is the case. The stronger the contingency, the higher were the corresponding judgements. Within each contingency, inspection of the slopes of the four curves indicates that participants' contingency estimates were lower for small samples than for the entire population of 40.

Insert Figure 2 approximately here

Statistical analyses support these observations. A 4 contingencies by 4 sample sizes by 2 pill colours repeated-measures analysis of variance yielded a main effect of contingency ($F(2.66, 77.08) = 82.36$, $MSE = 2129.16$, $\eta^2 = 0.74$), and a main effect of sample size ($F(2.71, 78.46) = 3.28$, $MSE = 2595.53$, $\eta^2 = 0.10$). No other effects were significant. Pairwise comparisons showed that estimates differed from each other across all contingencies (for ΔP = .20, mean judgement = 15; .40, 42; .60, 59; .80, 75), and that judgements for a sample size of 4 ($\overline{X} = 44$) were significantly lower than those based on the population of 40 ($\overline{X} = 54$). All the other pairwise comparisons for sample size were non-significant. This effect of sample size replicates Mercier and Parr (1996) but contradicts Kareev et al. (1997).

Since the entire population of trials was used in condition 40, the actual contingency (ΔP) did not vary in this condition whereas it did in conditions 4, 6, and 8. This does not confound sample size with sampling variability because the sample size sum of squares depends on deviations from the marginal mean for that effect while the sampling variability occurs within cell. Yet because the only significant difference was found between conditions 4 and 40, the fear is that the effect may be an artefact of the reduced within cell error variance in condition 40. To rule out this possibility, we obtained an unbiased estimate of the within cell error variance from a separate ANOVA designed with all the factors of the original analysis minus condition 40. The *MSE* for the sample size effect was 2363.40, actually lower than the 2595.53 in the whole model.

Dividing the sample size mean square from the whole model by the unbiased *MSE* from the reduced model, we therefore still obtained a significant effect for sample size ($F(2.71, 78.46) = 3.61$). Had the study been designed with condition 40 being a sample of yet a larger population, the problem would have remained because the variability of that large sample (40) would still be less than that of the smaller samples (4, 6 or 8). Apart from controlling this problem statistically at analysis time, the only other method would be to remove sampling fluctuations from all conditions by creating exact ΔPs for small numbers of trials. This is what Mercier & Parr (1996) had done and they also obtained lower judgements with smaller sample size.

Recall from the introduction that we chose to use sampling rather than fixed ΔPs in the current experiment to at least make it possible for the actual ΔP or an equivalent statistical index to be inflated since Kareev's narrow window approach implicitly requires that judgements be monotonically related to such a statistic. The observed ΔP values for the four target contingencies of .20, .40, .60 and .80, pooling over the three random sampling conditions of sizes 4, 6 and 8, averaged .19, .39, .61 and .82 respectively. This indicates that the decreased perceived contingency in the smaller samples did not depend on an objective property of the contingency between the stimuli. Thus the narrow window approach is doubly contradicted. First, because judgements were smaller in smaller samples; second, because whatever their direction, judgements are at least partially independent from statistical indices and their bias. The effect obtained by Kareev et al. (1997) seems to have reflected an artefact of the statistical bias inherent to the φ correlation that was performed by the researchers on participants' answers rather than contained in participants' judgements. The associative model, on the

other hand, correctly anticipated the discrepancy between judgements and statistical indices of covariation in the smaller samples irrespective of the sampling behaviour of the indices.

In defence of the narrow window approach, it could be argued that it was the same limited set of four contingencies that were presented repeatedly to the same participants while varying sample sizes. As people are excellent pattern detectors, they may have noticed that there appeared to be only four distinct values for the various relations and strategically used this information to shape their judgement. In other words, perhaps people were guessing the correct answer based on the overall pattern across conditions and, as a result, the overestimation in small samples forecasted by the narrow window approach did not occur. Be that as it may, the effect of the limited set cannot account for the fact that the results did not merely show a lack of larger estimates in small samples, they showed the opposite: larger estimates in larger samples, which the associative model can account for.

Another problem inherent to the narrow window approach is that, by postulating that people calculate contingencies using Pearson's correlation coefficient (or an equivalent), with small samples, this formula sometimes runs into indeterminate conditions. Indeed, as observed in Experiment 1, in extracting small samples randomly from the population of 40 discrete trials, it sometimes happened that the sample contained no event at all in cells *a* and *b*, or in cells *c* and *d*. In these cases, the normative ΔP is indeterminate because it involves a division by zero. This raises two issues. One, does indeterminacy cause problems for the participants or just for our mathematical method for computing the normative answer? Two, if indeterminacy is a source of difficulty,

what kind of judgements do the participants provide in these cases? Do they guess, thus making the data partially random, or do they answer zero as an equivalent to "I do not know", thus artificially lowering the average in the smaller samples where the likelihood of indeterminacy is higher?

Empirically, participants thought nothing of providing judgements under so-called "indeterminate" conditions. No participant ever complained of being asked the impossible. A closer inspection of the data revealed that there were only a few instances of indeterminacy. Averaging over contingency conditions and stimulus replications, 4 out of 30 participants encountered an indeterminate case for samples of size 4, 1 in 30 for samples of size 6, 0 in 30 for samples of size 8, and 0 in 30 for samples of size 40. The judgements for these points exclusively averaged .18 in the condition where sampling was from a .20 population, .38 when sampling from .40, .83 when sampling from .60, and .82 when sampling from .80. Thus, participants do not appear to guess and do not appear to fall back on answering zero. Nevertheless, in order to ensure that these few cases did not artificially lower the means, the data was re-analysed excluding all empirical judgements obtained under the "indeterminate" conditions, substituting the mean of all other participants in the same condition for these data points. The ANOVA results remained identical to the original analysis with significant main effects for contingency ($F(2.73, 79.15) = 86.89$, $MSE = 1939.23$, $\eta^2 = 0.75$) and sample size ($F(2.67, 77.43) = 4.29$, $MSE = 2505.43$, $\eta^2 = 0.13$). This aspect of the results is consistent with the associative model, as the model's equations never run into indeterminate cases. Indeterminacy is only a problem for the normative coefficients used by experimenters. It seems that the cognitive system has evolved an impressive algorithm that bypasses the

problems encountered by formal mathematics. This, however, raises an important issue for statistical models as well as for the narrow window approach, as they are both based on a calculation of a statistical coefficient (ΔP).

The fact that people could easily provide a judgement even when no information at all was available about one dimension of the problem (e.g., either cells *a* and *b* or cells *c* and *d* both equal 0) may seem to suggest that the data collected were not correlation judgements but something else. A correlation being a relation between two variables, both variables must exist before the strength of the relation can be assessed. Either that, or the value of the unknown variate must be taken as zero until evidence to the contrary becomes available. Indeed, it is plausible that for all problems, not just the indeterminate ones, participants provided answers that were not correlations in the statistical meaning of the term but were a reflection of *two* associative variables: the strength of the link between the presence of the medication and the skin discoloration outcome (Equations 3, 4), and that of the link between unidentified background cues remaining in the absence of the medication and the outcome (Equations 5, 6). These two associations are integrated in a single quantity via the term V_{Sum} in the equation system. Because V_{Sum} is subtracted from λ, the two associative values compete with one another such that the asymptotic value of A is given by Equation 11 (Cheng, 1997; Lober & Shanks, 2000),

$$V_{asymptote} = \frac{\beta_O a}{\beta_O a + \beta_{\bar{O}} b} - \frac{\beta_O c}{\beta_O c + \beta_{\bar{O}} d} \tag{11}$$

showing that, as long as the salience of trials with and without the outcome is equal ($\beta_O = \beta_{\bar{O}}$), the integrated associative strength is the same as ΔP. Thus, judgements based

on associative strength really are relational in nature; they do incorporate the two binary values (present / absent) on the two dimensions (cue / outcome) of the problem.

Degree of confidence in the contingency estimates was also analysed. The vertical positions of the curves in Figure 3 show that people were more confident in their answers when they were presented with a stronger contingency than a weaker one. Furthermore, the slopes of the confidence curves within contingencies show that participants were more confident when dealing with smaller samples as compared with larger samples.

Insert Figure 3 approximately here

Statistical tests once again support these assertions. A repeated measures ANOVA with the same design as for contingency estimates showed significant main effects of contingency ($F(2.99, 86.94) = 16.35$, $MSE = 355.81$, $\eta^2 = 0.36$), and sample size ($F(2.75, 79.76) = 6.44$, $MSE = 465.17$, $\eta^2 = 0.18$). Pairwise comparisons revealed significant differences in confidence between contingencies of .20 ($\bar{X} = 74$) versus both .60 (81) and .80 (85). Judgements of .40 (77) and .80 also differed reliably in confidence. Post-hoc comparisons of differences in confidence with sample size showed significant differences between confidence for samples of 4 (83) and 40 (75), and for samples of 6 (80) and 40. No other pairwise comparison was significant. Recall that Cheng and Holyoak (1995) postulated that contingency judgements and confidence were actually always confounded and that the learning curve reflected combined judgements (which remained unchanged with increasing sample size) with confidence (which increased with increasing sample size). However, it appears that contingency and confidence judgements vary

independently as a function of sample size. These results show that sample size not only influences confidence separately from contingency estimates, but that this influence on confidence estimates is even opposite to that on contingency judgements. Thus, the results obtained here are the only ones not open to the alternative interpretation of lowered confidence in smaller samples.

The randomization of sample size conditions also makes the results stronger than many of those previously obtained in a similar paradigm (e.g., Baker et al., 1993; Shanks, 1985; Shanks & Dickinson, 1987 – but not Mercier & Parr, 1996), where other conditions were randomized or counterbalanced but not sample size. In particular, the randomization of sample size here prevented any undesirable effect of multiple judgements to systematically bias the sample size data by spreading any such effect over all conditions with equal probability. At the same time, the randomization did not make it too hard for the participants to distinguish between samples as evidenced by the statistically significant effects.

Another strength of the study is that sample size was manipulated directly instead of relative to a measure of memory capacity such as the digit-span test (Kareev et al., 1997, Experiment 2) or instead of classifying participants according to a median split (Kareev et al., 1997, Experiment 1). A general weakness of these two alternative methods is that they are imperfect measures of memory capacity and may introduce unknown artefacts. For instance, the median split technique renders the definition of memory capacity dependent on the sample of participants recruited for a given study, which can make replications and comparisons across studies difficult. Assigning stimulus sample size relative to each participant's memory capacity is powerful in principle, but depends

on the particular aspect of working memory that is captured by the capacity test. For example, memory as measured by the digit-span test may not overlap exactly with the specific capacity of working memory that is taxed in covariation detection (Towse, Hitch, & Hutton, 2000). As explained in the introduction, associative theories imply that the memory for the sequence of events that the digit-span test quantifies may not be involved at all in contingency judgement. Manipulating sample size directly ensures that this stimulus dimension does interact with memory capacity. The present study is a case in point.

Asking participants to estimate covariations directly was also advantageous. Kareev et al. (1997) had chosen not to use such a technique on the grounds that it was "unnatural". They preferred to obtain trial-by-trial predictions and compute a contingency coefficient themselves from this partial information. The extent to which the direct collection of contingency estimates is unnatural is highly debatable considering that all research participants agree that they understand the task after practising it and before beginning the experimental trials. Also, trial-by-trial predictions merely require participants to track the relative frequencies of the outcome in the presence and in the absence of the cause, whereas judging contingencies requires comparing two such relative frequencies, making it a more elaborate cognitive process. To really assess covariation perception, the experimental task must ensure the collection of judgements that are relative in nature.

Inferring covariation perception from computations partially performed by the experimenter (e.g., φ based on probability estimates) is open to capturing biases of the computational method, which may be foreign to the cognitive process under study (e.g.,

coefficient φ may have biases that the cognitive system does not share if it does not compute φ or an equivalent). Some indirect indexes of covariation perception are free of this problem. Kareev et al.'s (1997) $\left|Z_{\bar{Y}} - Z_X\right|$ is an example. Nevertheless, both indirect indexes such as $\left|Z_{\bar{Y}} - Z_X\right|$ and direct estimates such as collected here are still not completely free of scaling difficulties. One problem is that it is not known where zero is on the mental scale. As a result, people may be quite good at discriminating the rank order of various contingencies even while they may appear somewhat inaccurate just because all of their judgements could be off the experimenter's response scale by a constant amount. It is also not known how the cognitive scale maps onto the experimental response. It is necessary to assume that the mapping is monotonic but the additional assumption of equality of intervals within the cognitive system is risky. For this reason, we really should only base our tests on comparisons of participants' estimates among themselves. Any attempt to compare participants' estimates with a normative measure is fraught with possible inconsistencies. This is probably why Kareev et al. found their low memory capacity participants to be, at the same time, "more extreme" and "at least as accurate as (and probably better than) participants with higher capacity" (p.282). If it is taken seriously, the accuracy of the low capacity group contradicts Kareev's theoretical prediction that low capacity engenders overestimation.

What is the theoretical impact of the current results? First, the reduced estimates produced by small samples are consistent with any associative analysis, in particular with Wagner's SOP (1981), and inconsistent with statistical models, including Kareev's (1995) narrow window approach. Associative models expected this result because the mechanisms of the associative process imply that less stimulus exposure entails less

strength build-up. Kareev, on the contrary, anticipated an increased objective and subjective covariation. One reason for the lack of consistency between the data and Kareev's (1995) narrow window approach, is that, as the results from Experiment 1 show, memory of the string of events leading to covariation perception is not required (provided that people don't have to reconstruct their covariation estimate). Thus, while it is true that working memory has limitations in terms of serial recall, this particular limitation does not happen to bear on the mechanism by which sensitivity to covariation is implemented in the cognitive system.

This is not to say, however, that short-term memory capacity has no influence at all on contingency judgements. Like other associative models, SOP assumes that associative strength cumulates over trials. In addition, the model details the real time process of strength accumulation on a single trial. According to SOP, the kind of limitation that impacts on the associative process is the number of stimulus elements that can be *simultaneously* active in short-term memory. Wagner (1981, p. 11) suggested that the limit might be in the order of approximately two or three stimuli. Other cognitive research suggests that the upper limit might be more in the order of seven (Miller, 1956). It is difficult to put an exact number on the ceiling capacity. Even if it could be done, there might be some performance deterioration before reaching the absolute limit. In contingency judgements tasks, stimulus trials are usually far enough apart to avoid reaching the limit on the simultaneous activation capacity. But when trials are massed together, more stimuli begin to need simultaneous activation in the real-time short-term memory process and the strength accumulation slows down due to mutual interference among stimuli. To explain why fast presentation of the stimulus sample of any size

reduced contingency estimates, Mercier and Parr called upon Wagner's SOP and this is how they interpreted their findings.

Although it was unclear which prediction Cheng's statistical model offered regarding contingency judgements and sample size, it was possible to derive two predictions: 1- an underestimation with small samples (comparatively to larger samples) represented by a reduced confidence intermixed with contingency judgement; 2- an overestimation with small samples due to a skewed distribution of statistical coefficients such as Pearson's correlation or ΔP (as proposed by the narrow window approach). Both predictions are contradicted by the current results. First, if we assume that the bias raised by Kareev should influence the calculation of ΔP (if people indeed calculate this statistic or an equivalent), then her theory should also predict an overestimation with smaller samples and, as was shown, this is not what was obtained. Second, if we suppose that judgements remain constant across sample size but that confidence increases with sample size then, her theory could, at first sight, seem to account for what was observed.

Indeed, until now, one limitation of interpreting contingency estimate reductions in associative terms was that the prediction of decrements in associative strength was confounded with potential conservatism in the estimation process when the data are obtained under a less than ideal sampling procedure. That is, people might conservatively give answers closer to zero when they are less confident in their estimation, as a means to say, "I do not know", as proposed by statistical models. Mercier and Parr's accelerated stimulus presentation rate is one case in point as, unfortunately, they did not measure confidence. The current study might also be conceived as having created less than ideal sampling when the samples were small. Interestingly, however, confidence was *higher*

instead of lower with the small samples. This rules out any alternative explanation of the current results in terms of reduced confidence for small samples and contradicts the statistical models' potential account for the observed results. It also demonstrates that confidence and associative strength are separable psychological processes. Finally, an intuitive explanation of why confidence is higher in smaller samples could be that, in general, people are subjectively aware of their limitations in terms of memorizing a string of episodic events. Thus, when they are asked to report confidence, people rate it higher if the sample size is smaller, feeling that they can handle the task better, unaware that the task does not really tap into the memory capacity in question.

While confidence was inversely proportional to sample size, it was also directly proportional to the normative contingency value; that is, the smaller the ΔP, the lower the confidence. Interestingly, for ΔP closer to zero, frequencies in the four cells of the contingency table tend to be more equally distributed. Thus, it is possible that this aspect of confidence is related to the diversity of the objective experience when ΔP is small.

The main goal of this study was to determine what happens to contingency judgements when based on small samples of episodic events. Kareev's (1995) narrow window approach predicted that judgements would be higher while Wagner's SOP predicted that judgements would be lower. The results contradicted Kareev's viewpoint in favour of Wagner's and the associative approach in general. When considered in the light of many other experiments (see Shanks, Medin, & Holyoak, 1996) and the current results, the associative approach appears to have greater power and generality. Statistical models of how the mind works often imply that our cognitive system proceeds in a normative and rational fashion. But it is becoming increasingly clearer that the cognitive

system does not function in such a manner. Instead, the cognitive system uses brain computational mechanism selected in an evolutionary manner for their adaptive value. The beauty of this view is that the associative mechanism is a plausible candidate for such evolutionary selection in that it explains covariation sensitivity in both contrived laboratory tasks and in naturally occurring behavioural contingencies.

Experiment 2

Results from Experiment 1 confirm Wagner's SOP theory and contradict Kareev's narrow window approach. When contingency judgements were based on smaller samples, they were smaller as opposed to larger than the normative population value. Wagner's associative approach anticipated this result not because of a limitation in memory capacity but because small samples of trials yield less opportunity for associative strength build-up. However, these results should not be interpreted to mean that limitations in memory capacity are completely denied or that any such existing limitation has no impact on associative systems, quite the contrary.

In accordance with memory research, Wagner does postulate some limitations on memory capacity. However, recall from the introduction that the limitation that would impact on the associative process, according to Wagner, is the number of stimuli that can be *simultaneously* active in memory and not the length of the series of individual events. Remember that Wagner postulates severe limitations on the *A1* state, such as only two or three nodes could be *fully* in this state simultaneously. As stated previously, the summed proportion of elements that can be simultaneously in their *A1* state across all memorial nodes is represented by a constant (*C1*). In order to keep the proportion of elements in their *A1* state to *C1*, as the number of activated nodes increase in memory, this

momentary increase must be "compensated for by some decrease in the proportion of *A1* elements over all nodes so that the summed proportion will still equal *C1*" (Wagner, 1981, p. 23). Although, up to ten or fifteen nodes could be in an *A2* state simultaneously, according to Wagner, the link between nodes cannot be modified while elements are simultaneously in this state. Following this set of assumptions, since a proportion of the elements in *A1* state will decay faster in the presence of multiple simultaneous contingencies, the link formed between these nodes should be correspondingly weaker (as this quick decay will prevent the normal growth of the association). We should therefore observe an underestimation of the contingency between these stimuli as the number of simultaneous contingencies increases (as compared to a situation where memory load is not exceeded). Experiment 2 is aimed at testing this prediction.

It is important to note that this experiment does not use the design usually observed in overshadowing studies, where participants are asked to judge the effect of multiple causes presented simultaneously with a single outcome. Rather, multiple causes as well as multiple outcomes were used. All contingencies were separate, in that, for each outcome, there was always only a single cause. Therefore, when there were two or three medications to evaluate, there were also two or three separate different contingencies to assess with their separate putative cause (i.e., medication) and outcome (i.e., skin colour).

Method

Participants. Participants were 24 volunteer students (19 females and 5 males) from the University of Ottawa (21 undergraduate and 3 graduate students), aged between 18 and 45 years ($\overline{X} = 26$).

Apparatus. The experimental task was administered using an IBM compatible microcomputer, equipped with a 14 inch VGA monitor, which was also used to collect the data. Programming of the study was done using Microsoft Visual Basic Professional 6.0.

Stimuli. Through sequential trials, participants were shown fictitious medical files of different patients suffering from a disease inducing skin discoloration. The medical files were represented using three 10 cm by 15 cm frames, placed one above the other and centered on the computer screen. These frames were always present, regardless of the experimental series. When less than three contingencies were presented simultaneously, the extra frames simply remained empty. The target contingency always appeared in the top window, the first additional contingency in the second window and the second additional contingency appeared in the bottom window. Each file depicted either the presence or the absence of treatment and the presence or the absence of skin discoloration. The treatment (present or absent) appeared on the left portion and its effect (present or absent) was presented on the right. Presence of treatment was symbolized with red, blue or green oval-shaped medication pills measuring approximately 9 mm in height and 5 mm in width. To represent absence of medication, the same oval-shaped pills were used with a black "X" overlaid on the pill's image. Presence of skin discoloration was represented using a round icon of a sick face (tongue sticking out) in the colour of its corresponding treatment (blue, red or green). Absence of skin discoloration was shown using a round icon of a white smiling face. The face icons measured approximately 12mm in diameter. For trials containing one, two or three

medical files, the files remained visible for 1, 2 or 3 s respectively, with 1s inter-trial intervals. Examples of the stimuli used are shown in Figure 4.

Insert Figure 4 approximately here

Procedure. Upon their arrival at the laboratory, participants were led to an individual testing room where they were informed of the experimental procedures on computer screen. The experimenter read the following directions:

> Welcome to the Cognitive Psychology Laboratory. We are studying decision-making processes such as those used by health professionals. Imagine that you have access to the medical files of many people all sick with certain diseases. These diseases have many symptoms, including a skin discoloration. Different medications treating these diseases have been used in clinical trials. The problem is that the medications themselves can also increase or decrease the risk of skin discoloration. We will ask you to evaluate different medications and different diseases. Depending on the case, you may have to judge the efficacy of one, two or three different medications simultaneously [an example of each combination was displayed on the screen as it was mentioned in the text of the instructions]. For each medication, the clinical trial contains the results of many individual files. Each file may report absence of medication or presence of medication, the presence of skin discoloration or the absence of skin discoloration, in different possible combinations. You will have an opportunity to warm up your diagnostic

skills with three practice trials. The first practice trial will use one medication and one disease, the second, two, whereas the last one will use three.

These instructions were followed by three series of practice trials. A total of 40 medical files were shown for each contingency involved in the practice trials. The first practice had a contingency ΔP of .40. Using a 2 x 2 contingency table where cell *a* represents the frequency of the presence of both events, cell *b* represents the frequency of the presence of the treatment but the absence of skin discoloration, cell *c* represents the frequency of the absence of treatment but the presence of skin discoloration, and cell *d* the frequency of the absence of both events, the frequencies for the first practice series were 14, 6, 6 and 14 for cells *a*, *b*, *c* and *d* respectively. Trial order was randomized. Practice two contained two different contingencies presented simultaneously. The first contingency used the same ΔP .40 based on the same frequencies of 14, 6, 6 and 14 as in practice one with a new trial randomisation. The same frequencies were re-used to avoid any bias due to differing frequencies within contingency. The second contingency was a ΔP of -.40 with frequencies of 6, 14, 14 and 6. The third practice set presented three simultaneous contingencies. Two of these contingencies used ΔPs of .40 along with a null contingency with frequencies of 10, 10, 10 and 10. For the practices only, the participants were informed of the true value of each pill's contingency at the beginning of the practice series. The same contingency of .40 was used repeatedly to ensure that participants focused their attention on the contingency that differed.

After each practice, participants were asked to judge the efficacy of the treatment relative to the absence of the treatment based on all files presented for the clinical trial and for each medication presented (one, two or three). Participants gave their answer

using a scale that ranged from -100 to 100, where -100 meant "extreme reduction of the risk of skin discoloration", 100 meant "extreme increase of the risk of skin discoloration", and 0 meant "absence of change in the risk of skin discoloration". The experimenter was present for the practice trials in order to answer questions. No direction was given as to suggest any one particular way of arriving at an answer. Participants were told that they might have the impression of sometimes guessing the answers but that this impression was normal. They were also told that the test series could contain one, two or three different medications presented simultaneously and were asked to try to always evaluate each medication as best as they could. It was also emphasized that the relations between treatments and effects were constant within a clinical trial but varied between clinical trials. Data from the practice series were omitted from the analyses.

After ensuring clear understanding of the task, the experimenter left the room and participants began the test series. A total of 14 series of clinical trials were presented randomly to the participants. These test series used target contingencies of ΔP .20 and .80 with frequencies for cells *a*, *b*, *c* and *d* of 12, 8, 8 and 12 and 18, 2, 2 and 18 respectively. Each of these target contingencies was presented alone or along with all combinations of either one or two additional contingencies (for totals of one, two or three simultaneous contingencies respectively) using a ΔP of .20 and .80 (with the same cell frequencies as the target contingencies). The designations of target and additional contingencies are used solely for the purpose of clarifying the procedure; the participants were unaware of these designations. The design of Experiment 2 is shown in Table 1.

Insert Table 1 approximately here

Colour of pill (blue, red or green) was attributed randomly with the constraint that each pill within a series received a different colour. At the end of each series of medical files (40 clinical trials containing one, two or three medical files each), the judgement screen appeared with the rating scales. The pills were presented in the same top to bottom position on the screen as they occupied during the display of series. The contingency judgement scale was reset to 0 (middle of the scale) for each new judgement screen. Participants could change their answers before moving on to the next clinical trial but they could not return to their current judgement screen after submitting their responses. All medications had to be evaluated before moving on to the next trial. Before each new series, a warning window appeared reminding participants to start their evaluation afresh as they were about to start a new series of clinical trials.

Results and discussion

All data were analysed using repeated-measures analyses of variance (ANOVA) with a Type I error rate of .05. To protect against possible violations of sphericity assumptions, a Huynh–Feldt correction was applied to all repeated measures with >1 df in the numerator. Tukey's honestly significant difference (HSD) test was used for post-hoc pairwise comparisons.

The main results for contingency estimates of the target contingency are presented in Figure 4. In that figure, values for the target contingencies were averaged over replications in conditions where multiple contingencies occurred. For instance, in the

double contingency condition, target ΔP = .20 occurred twice: once accompanied with a second contingency of the same value (.20, .20) and the second time accompanied with a second contingency of .80 (.20, .80). In order to validate the experimental task, it must first be ascertained that target contingencies of .20 and .80 were well discriminated. The slopes of the three lines in Figure 5 suggest that this was the case. The distance between the three lines also indicates that adding a second contingency had the effect of lowering judgements, but that adding a third contingency did not cause further deterioration in judgements.

Insert Figure 5 approximately here

Statistical analyses support these assertions. A number of simultaneous contingencies (3 levels: 1, 2 or 3 medications) by contingency values (2 levels: ΔP = .20 or .80) repeated-measures analysis of variance on target contingencies yielded a main effect of number of simultaneous contingencies (F (1.87, 42.91) = 11.83, MSE = 507.91, η^2 = 0.34), and a main effect of contingency value (F (1, 23) = 186.81, MSE = 479.24, η^2 = 0.89). No other effects were significant. The main effect of contingencies demonstrates that contingencies of .20 ($\overline{X}$ = 14.91) and .80 ($\overline{X}$ = 64.77) were well discriminated across conditions. Pairwise comparisons demonstrated that contingency estimates were significantly higher when participants were presented with one contingency ($\overline{X}$ = 52.31) as opposed to two ($\overline{X}$ = 33.29) or three simultaneous contingencies ($\overline{X}$ = 33.92).

These results support Wagner's SOP theory. Wagner postulated that only 2 or 3 nodes could be fully in their *A1* state simultaneously. When a single contingency was

present, a total of three nodes were simultaneously activated (since each stimulus – treatment and effect – as well as the context, activate their own respective node). This is within the assumed capacity limit and therefore judgements are free to reach asymptote. However, as soon as a second contingency is introduced, the number of nodes simultaneously in their *A1* states totals five (one for the context and each of the four stimuli), which goes beyond the assumed limit. As stated previously, when the limit is surpassed, the decay process is accelerated and, as a result, asymptote cannot be fully reached within the same number of trials. Consequently, judgements should be underestimated in these cases (comparatively to cases where memory capacity is not taxed). This is exactly what was obtained in the current experiment. Participants' judgements were underestimated when participants were presented with two or three simultaneous contingencies (for a total of up to seven nodes simultaneously activated).

A further analysis was performed on the subset of data where participants evaluated three simultaneous contingencies. This analysis has three interesting features. First, unlike the previous global analysis, the data points were not averaged across conditions, thus ensuring equal likelihood of variability across cells in the design. Second, it investigates whether the effect of the added contingencies is constant across differing values of the additional judgements. Third, it tests subjective estimates of the additional contingencies, as well as the target judgements. A 3 types of stimulus (target, first additional contingency or second additional contingency) by 2 values of the target contingency (ΔP = .20 or .80) by 2 values of the first additional contingency (ΔP = .20 or .80) by 2 values of the second additional contingency (ΔP = .20 or .80) repeated-measure analysis of variance was performed. This yielded main effects of target contingency (*F*

(1, 23) = 31.25, *MSE* = 1453.92, η^2 = 0.58), first additional contingency (*F* (1, 23) = 16.55, *MSE* = 1561.26, η^2 = 0.42), and second additional contingency (*F* (1, 23) = 32.43, *MSE* = 1227.44, η^2 = 0.59), confirming that each contingency was well discriminated in its own right ($\overline{X}$ = 22.88, 25.07 and 23.45 for the target contingency, the first additional contingency and the second additional contingency respectively when ΔP = .20 and $\overline{X}$ = 40.65, 38.46 and 40.08 when ΔP =.80). The three panels of Figure 6 respectively correspond to the significant interactions of type of stimulus (i.e., each medication) by target contingency (i.e., the judgement of all types of stimuli – target, first additional contingency, second additional contingency – when the value of the target contingency presented with them was .20 and .80) (*F* (1.59, 36.53) = 34.74, *MSE* = 933.35, η^2 = 0.60), type of stimulus by first additional contingency (i.e., the judgement of all types of stimuli when the value of first additional contingency presented with them was .20 and .80) (*F* (2, 46) = 31.54), *MSE* = 891.77, η^2 = 0.58), and type of stimulus by second additional contingency (i.e., the judgement of all types of stimuli when the value of second additional contingency presented with them was .20 and .80) (*F* (1.83, 41.98) = 28.90, *MSE* = 1088.58, η^2 = 0.56). This confirms that the overall effect of multiple contingencies described as the main result of the experiment in Figure 5 is reproduced within this subset of the data.

Insert Figure 6 approximately here

In addition, there were significant interactions of target contingency by first additional contingency (*F* (1, 23) = 4.86, *MSE* = 1552.83, η^2 = 0.17) and first additional

contingency by second additional contingency (F (1, 23) = 7.30, MSE = 710.29, η^2 = 0.24). No other effects were significant. The interactions between the target contingency and the first additional contingency and between the first additional contingency and the second one are shown in Figure 7.

Insert Figure 7 approximately here

The top panel of Figure 7 indicates that the target contingency was judged lower when presented alongside a first additional contingency of .20 than when presented with a first additional contingency of .80. This downward effect was stronger for a large value of the target (.80) than for a small one (.20), perhaps simply because of floor effect. A similar situation arises in the middle panel relating judgements of the first additional contingency with the nominal values of the second additional one. However, there was no such relation between judgements of the target and the nominal values of the second additional contingency (bottom panel).

The influence of the contingencies on each other is very interesting. In general, the presence of two strong contingencies seems to result in higher contingency judgements. Similarly, two low contingencies seem to lead to lower estimates and a high and low contingencies presented simultaneously seem to give intermediate estimates. Since this effect was only present between contingencies that were presented adjacent to one another in the judgement display, it is easy to view these results in terms of contextual effects. Recall that Wagner (2003) postulated, through his replaced elements conception, that a portion of the elements from the context would contribute to the

formation of associative strength of a given contingency. Since the portion of elements from stronger or weaker contingencies contributed to the formation of associative strength of the contingency, it is easy to see why the estimated value of these contingencies would be higher when presented with two strong contingencies (stronger context) and lower when presented with two weaker contingencies (weaker context).

Experiment 3

Although Experiment 2 yielded interesting results, they were not as definite as can be. First and foremost is the fact that an alternative explanation exists to account for the main results. Even though degradation in contingency estimates with increasing number of contingencies was observed, it could both be explained by a decrease in associative strength as postulated by Wagner or a decrease in confidence. Indeed, it is possible that when facing multiple contingencies, participants were less confident in their judgements and therefore were also more conservative in their answers giving estimates closer to zero. Because confidence was not assessed concurrently, it is impossible to determine which explanation holds. To counter this problem, confidence was measured after each judgement in Experiment 3.

Furthermore, it was not clear if the total capacity limit was reached in Experiment 2. In accordance with Wagner's assumptions, when the limit is exceeded, the proportion of elements in their primary state of activation should decrease as the number of contingencies presented simultaneously increases, therefore resulting in weaker associative strength. In Experiment 2, no further decrease in judgements was observed beyond those caused by the addition of one extra contingency. Yet there was ample room for further judgement reduction especially when $\Delta P = .80$ with three simultaneous

contingencies. Why did the participants keep discriminating so well between $\Delta P = .20$ and $\Delta P = .80$, even with three simultaneous contingencies to track? Two aspects of the task might have made it unusually easy for the participants to maintain a high level of discrimination. First, only two contingency values, .20 and .80, were used throughout the experiment. Second, these values were very distinct from one another. It is probable that participants noted the pattern after a few series of clinical trials and adjusted their judgements accordingly so that although two or three simultaneous contingencies caused a decrease in judgements, they might have still been quite easy to discriminate. The pattern might have been easier to detect as participants were expressly told that, in the practice trials, the same contingencies were repeated.

In order to further test the memory capacity, degree of difficulty was increased in Experiment 3. The target contingencies selected were closer in range to one another (.40 and .60), making them harder to discriminate. The choice of .40 instead of .20 as the low value for the target contingency also remedied another difficulty in interpreting results from Experiment 2 where there was a possible floor effect. Additional contingencies were different from the values of the target contingencies (0, .20, .80 and 1) for a total of six different values throughout the experimental task. The maximum number of simultaneous contingencies was increased to four. Furthermore, although some contingencies were repeated in the practice trials, participants were not made aware of this fact. Instead of indicating the exact value of the medications to the participants during practice, the instructions only informed them of whether the contingencies were positive or negative as well as strong or weak.

Method

Participants. Participants were 24 volunteer undergraduate students (15 females and 9 males) from the University of Ottawa, aged between 20 and 37 years ($\overline{X} = 26$).

Apparatus and stimuli. The apparatus was identical to that used in Experiment 2. Only slight changes were introduced in the stimuli. First, an extra frame was added to the stimulus display so that 4 medications could be visible at the same time. Series showed one, two or four simultaneous medical files. Position of treatment and outcome for a given contingency was randomly selected so that when only one contingency was present, it could be located in any of the four frames, the others remaining empty. When two contingencies were presented, they were located one immediately on top of the other either in the top, middle or bottom two panels, the others remaining empty. When four contingencies were presented, the position of each was determined randomly. This random assignment of contingencies was made across series and across participants but not within series. Medical files and presence or absence of treatment or outcome were identical to Experiment 2 with the exception that an additional purple medication was added in order to have four different stimuli.

Procedure. The same directions were given to the participants on the computer screen as those shown in Experiment 2, with the difference that they announced that participants would be asked to judge the efficacy of one, two or four simultaneous medications instead of one, two and three. Similarly, they were told that the three practice trials would contain one, two and four contingencies. The instructions were followed by three series of practice trials. The first practice used a single contingency of .70 with frequencies of 20, 0, 6 and 14. The trial order was randomized. Practice two had two

simultaneous contingencies of -.70 and .70. The frequencies were 0, 20, 14 and 6 for the contingency of -.70. The last practice used four medications with contingencies of -.70, -.30, .30 and .70. The frequencies for the contingencies of -.30 and .30, were 4, 10, 16, 10, and 16, 10, 4, 10, respectively. A total of 40 medical files for each contingency were shown to the participants. At the onset of each practice, a window appeared informing participants of the upcoming series. For the first practice, the window text reminded participants that they would see only one medication and that it would increase the risk of skin discoloration. For the second practice, participants were reminded that they were about to see two different medications: one would increase the risk of skin discoloration and the other one would decrease it. In the window announcing practice three, participants were warned that they were about to view four simultaneous medications. They were told that two of these would increase the risk of skin discoloration; one more than the other, and that the two other medications would reduce the risk of skin discoloration and again, one more than the other. After each practice trial, participants judged the efficacy of each treatment relative to the absence of the same treatment based on all files presented for the trial, using the judgement scale described in Experiment 2. Stimuli appeared in the same order that they appeared during the trial. After giving their judgements, participants were asked to assess the degree of confidence in their judgements using a scale ranging from 0 to 100, where 0 meant "no confidence", 50 meant "moderate confidence" and 100 meant "extreme confidence". After the second and third practice trials, a feedback window appeared giving participants the correct answer (which contingencies were positive and which were negative). Data from the practice trials were once again omitted from the analyses. The experimenter was present for the

duration of the practice series to read the instructions and answer questions. The same directions as in Experiment 2 were used, except that no indication on the exact values of the contingencies was given and participants were not told that the same contingencies were repeated.

Following the practice series, 10 series of 40 trials each were presented in random order of series and trial within series. As shown in Table 2, the series were designed such that participants would always judge target contingencies of .40 (with frequencies of 14, 6, 6 and 14) or .60 (16, 4, 4, 16). In condition 1, these targets were presented separately and singly. In condition 2, the target contingencies were accompanied with one additional contingency, the value of which was either .20 below or .20 above the target. In condition 3, each target contingency was accompanied by three additional contingencies. The values of the additional contingencies were distributed with either two below and one above or one below and two above the target. When the contingency was 0, the trial frequencies were 10, 10, 10 and 10. When it was .20, the trial frequencies were 12, 8, 8 and 12; .80: 18, 2, 2 and 18; 1: 20, 0, 0 and 20.

Insert Table 2 approximately here

After each series of medical files, the judgement screen appeared and was followed by the confidence scale. Scales were reset to the middle. Participants were able to change their answers before passing to the next screen or starting the next clinical trial, but they were not able to return to previous screens. Before each new clinical trial, a window reminded the participants to start their evaluation afresh as they were about to

start the evaluation of a new clinical trial. No indication was given prior to each test trial on the values of the contingencies, nor was feedback presented afterwards.

Results and discussion

All data were analysed using repeated-measures analyses of variance (ANOVA) with a Type I error rate of .05. To protect against possible violations of sphericity assumptions, a Huynh–Feldt correction was applied to all repeated measures with >1 df in the numerator. Tukey's honestly significant difference (HSD) test was used for post-hoc pairwise comparisons.

The main results for contingency judgements of target contingencies are depicted in Figure 8. As in Experiment 2, the slopes of the lines suggest that target contingencies were well discriminated across conditions. The position of the lines suggests that contingency estimates were higher when only one contingency was present than when two or four contingencies were presented.

Insert Figure 8 approximately here

Statistical analyses confirm this. A 3 number of simultaneous contingencies (1, 2 or 4 contingencies) by 2 target contingencies (.40 or .60) repeated-measures analysis of variance revealed a main effect of number of simultaneous contingencies (F (1.99, 45.86) = 10.31, MSE = 943.59, η^2 = 0.31) and a main effect of target contingency (F (1, 23) = 14.14, MSE = 781.84, η^2 = 0.38). No other results were significant. The main effect of target contingencies show that contingencies of .40 ($\overline{X}$ = 29.59) and .60 ($\overline{X}$ = 47.12) were well discriminated across all conditions. Pairwise comparisons revealed that

estimates of target contingencies presented alone ($\overline{X}$ = 54.13) were significantly higher than when presented with an additional contingency ($\overline{X}$ = 34.41) or three additional contingencies ($\overline{X}$ = 26.53). No significant difference was found between the presentations of two versus four simultaneous contingencies.

These results replicated those obtained in Experiment 2. Once again, they support Wagner's SOP theory and the predictions concerning the limitations of the number of elements fully in their *A1* state simultaneously. Even though the difficulty of the task was increased, a further decrease in judgements between two and four contingencies still failed to reach significance, revealing that people might be even better at judging simultaneous contingencies than we assumed.

A second set of analyses was conducted in order to study the impact of the value of the additional contingencies. A 4 values of additional contingencies by 2 target contingencies repeated-measures analysis of variance showed a marginally significant discrimination of the target contingencies (F (1, 23) = 4.28, MSE = 898.38, p = .05, η^2 = 0.16) and a main effect of the value of additional contingencies (F (3, 69) = 3.12, MSE = 1944.82, η^2 = 0.12). No other effect was significant. As the position of contingencies was free to vary and because different additional contingencies were used, it was not possible to investigate the effect of adjacent contingencies specifically. However, planned comparisons show that when lower additional contingencies were used (0 and .20), target estimates were lower than with all other combinations of higher combined value (.20 and .80 or .80 and 1) (F (1, 23) = 37.19, MSE = 2284.46), as depicted in Figure 9. This replicates the contextual effects obtained in Experiment 2. This effect will be revisited in the General Discussion.

Insert Figure 9 approximately here

Confidence estimates were also analysed. The main effect for estimates of confidence in judging one, two and four simultaneous contingencies is shown in Figure 10. The position on the lines on the graph suggests that participants were more confident in judging one contingency than judging two or four contingencies presented simultaneously. The slopes suggest that participants were as confident in judging a contingency of .40 as one of .60, except perhaps in the case where two contingencies were presented simultaneously.

Insert Figure 10 approximately here

Statistical analyses support these observations. A 3 number of simultaneous contingencies by 2 target contingencies repeated-measures analysis of variance revealed a main effect of number of simultaneous contingencies (F (2, 46) = 47.86, MSE = 229.57, η^2 = 0.68) and an interaction of number of contingencies by value of target contingency (F (1.5, 34.51) = 4.41, MSE = 123.33, η^2 = 0.16). Pairwise comparisons confirm that participants were more confident in judging one contingency ($\overline{X}$ = 82.85) than judging two ($\overline{X}$ = 65.46) or four simultaneous contingencies ($\overline{X}$ = 52.71). Participants were also significantly more confident in judging two rather than four simultaneous contingencies. Participants showed no difference in confidence in judging target contingencies of .40 and .60, except when judging two simultaneous contingencies, where they were more

confident in judging a contingency of .60 ($\overline{X}$ = 69.94) rather than judging a contingency of .40 ($\overline{X}$ = 60.97). In Experiment 2, results could both be explained by reduced associative strength or decrease in confidence. Results from the current experiment indicate that a decrease in confidence cannot fully account for the results because the pattern of differences in confidence differs from the one observed with contingency estimates. Participants' confidence continued to decrease with increasing number of simultaneous contingencies even though contingency judgement performance was maintained from two to four contingencies. This suggests that confidence may be related more directly to the perceived difficulty of the task rather than actual judgement accuracy.

It is unclear why participants felt more confident about their .60 than their .40 judgements only when these targets were presented along with one additional contingency, either below or above the target in value. Recall that results from Experiment 1 indicate that confidence increases with the value of ΔP. Similarly, in this experiment, confidence in judging the target contingencies reliably increased with the ΔP of the additional contingencies as shown in Figure 11 and confirmed by a 4 values of additional contingencies by 2 values of target contingencies repeated-measure analysis of variance where the main effect of the value of additional contingencies (F (2.89, 66.54) = 4.08, MSE = 751.55, η^2 = 0.15) was the only significant factor. On that basis alone, we should have observed the difference in confidence between the judgements of .40 and .60 at all levels of number of simultaneous contingencies. However, it remains possible that confidence is sensitive to many factors that are not yet fully understood. For instance, it may be that confidence was equally high for .40 and .60 when they were presented alone

because the task was extremely simple then. Conversely, confidence may be equally low across target ΔPs when they are accompanied with three other contingencies because the overall level of difficulty is high enough to mask other effects. Pairwise comparisons on the main effect of the values of additional contingencies reveal that confidence in judging target contingencies was significantly lower ($\overline{X}$ = 43.42) when they were accompanied with additional contingencies of 0 and .20 rather than contingencies of .80 and 1 ($\overline{X}$ = 62.25). There were no other significant comparisons.

Insert Figure 11 approximately here

In accordance with Wagner's SOP theory, memory load is taxed by increasing the number of nodes that are simultaneously in their active states. Wagner postulated that only 2 or 3 nodes could be *fully* in their *A1* states simultaneously. When this limit is surpassed, a proportion of the elements in *A1* state would be forced to decay in their *A2* state more quickly (before associative strength can be fully built). This would prevent asymptote from being reached and contingency estimates should correspondingly be lower. Experiment 2 and 3 empirically verified this prediction. Results from both experiments fully support Wagner's hypothesis. In both cases, as soon as the number of nodes simultaneously in their *A1* state was increased to 5 (by presenting two simultaneous contingencies: 2 cues, 2 consequences and one node for the context), contingency estimates were significantly lower than when a single contingency was presented (3 nodes – cue, consequence and context – in their *A1* states).

An implicit expectation of the model is that memory capacity and hence, perceived contingency through associative strength, should further decrease as the pressure on the system grows. Surprisingly, presenting three or four simultaneous contingencies, for a total of up to 7 and 9 nodes simultaneously in their *A1* states, did not result in further performance decrements. Participants performed in these conditions as well as in the condition where two contingencies were presented simultaneously and they were still able to discriminate between contingencies, even close ones like .40 and .60 in Experiment 3. Although this finding does not contradict the memory capacity model per se, it leaves us wondering why judgements did not decrease further. Note that the SOP theory is not very explicit on how many elements are in a node. It is possible that the memory system can only compensate for the shortage of nodes by reducing the number of elements in their primary activation state only up to the equivalent of about four nodes. Beyond this point, different coping strategies may need to be envisaged. If the internal state of the system cannot be adjusted beyond the equivalent of four nodes, then further compensation could occur earlier in the chain of events that is, at the level of the sensory input. Since there are 40 trials supporting each contingency, participants may begin to sample the trials. Judgements based on samples are also expected to yield less associative strength (see results from Experiment 1; Mercier & Parr, 1996) but that reduction would affect all contingencies equally thus preserving their rank order discrimination. This hypothesis could be tested in future experiments by combining the sample size and the number of contingencies manipulations.

Another potential explanation for the lack of further decrement can be found in studies on retrospective revaluation (where people subsequently modify a prior

evaluation of a contingency). Indeed, in order for SOP to account for retrospective revaluation, Dickinson and Burke (1996) proposed that excitatory connections may be formed not only when nodes are simultaneously in their *A1* state but also when they are simultaneously in their *A2* state. This could perhaps be used to explain why participants' judgements did not decrease any further, even with four simultaneous contingencies. Recall that Wagner hypothesized that as many as 10-15 nodes could be fully activated simultaneously in *A2* state. Therefore, even in cases where four simultaneous contingencies were used (for a total of 9 nodes), this would still be within the limits of the *A2* state's capacity. It is also possible that, although links can be formed between elements simultaneously in their *A2* states, that such links are also weaker (as they are on their way to return to their inactive state). This could explain why, if links can be formed in *A2* state, they did not reach the level obtained with the presentation of a single contingency. Although this may seem to suggest that the distinction between *A1* and *A2* states would then become unnecessary, it is important to keep in mind that this is not the case. Although excitatory links may be built either when elements are in their *A1* or *A2* states simultaneously, inhibitory links are still being built when elements from one node are in their *A1* state while the others are in their *A2* state. For the time being, however, this suggestion remains speculative and will require further empirical testing.

While in Experiment 2 lower contingency estimates in conditions in which two or three contingencies were presented simultaneously could also be accounted for by a corresponding decrease in confidence, this alternative explanation was discounted in Experiment 3. When confidence was measured separately, results showed that it does not entirely covary with the contingency judgements. Contingency estimates sometimes

changed with the number of simultaneous contingencies and sometimes not, whereas confidence always changed with the number of simultaneous contingencies. Additionally, confidence sometimes changed with ΔP values and sometimes not, while contingency estimates always did. These results indicate that confidence is clearly a separate process with its own set of determinants that overlaps only partially with those of associative strength and contingency judgements.

Experiment 4

The results from experiments 2 and 3 are consistent with Wagner's view. The presentation of multiple contingencies was used to show that it is the simultaneity of memory nodes activation that is the crucial memory limitation in contingency judgements. The first experiment has shown that curtailing the amount of rehearsal at the level of a single pair of memory nodes would disrupt the accumulation of associative strength and hence reduce subjective perceptions of contingency. Experiments 2 and 3 add that contingency judgements are reduced by the simultaneous activation of multiple nodes.

Another aspect of limitations in memory capacity that deserves closer examination is that of individual differences. Wagner hypothesizes a difference in contingency estimates between high and low memory capacity groups when multiple contingencies are presented simultaneously as, in his view, this is when the memory capacity is taxed. Indeed, it is possible that the constant $C1$ for people with a lower memory capacity is lower than for those who possess a higher memory capacity. Experiment 4 is aimed at verifying that when multiple contingencies are presented

simultaneously, individuals with lower memory capacity will show greater deterioration in judgements than will individuals with higher memory capacity.

As stated in the discussion of Experiment 1, memory as measured by the digit-span test may not overlap exactly with the specific capacity of working memory that is taxed in covariation detection, as memory for the sequence of events that the digit-span test quantifies may not be involved at all in contingency judgement (see results of Experiment 1). However, although memory capacity involved in contingency judgment may not be in the form that Kareev (1995) suggests, it is still possible that any measure of working memory correlates with the memory capacity taxed in contingency judgements and that is why Kareev et al. (1997) obtained a significant memory effect. For this reason, as well as to investigate Kareev et al.'s effects, the digit-span test will be used in Experiment 4 and 5 to divide the participants into high and low memory capacity groups. Although Kareev et al. (1997) only used a forward digit-span test, Experiment 4 and 5 will use total scores on both forward and backward digit-span tests. In the backwards digit-span test, participants are asked to recall the sequence in the reverse order than it is presented. This task measures more than simply the recall of sequential events as it requires some "work" to be done on the information as it is re-ordered. This additional test should begin to tap into processing efficiency in addition to capacity defined purely in terms of number of available memory slots.

Also, the study by Kareev et al. (1997, Experiment 1) was criticized for the use of the median-split to divide the high and low memory capacity groups. This technique renders the definition of memory capacity dependent on the sample of participants for any given study. In order to reduce this problem, rather than the median split, out of a

group of 48 participants who participated in Experiment 4, it was decided to retain the 15 participants with the highest scores and the 15 participants with the lowest scores, to be included in the high memory capacity group and low memory capacity group, respectively. The 15 participants making up the high memory capacity group had obtained a total score on the digit span test (forward and backward) of 20 or higher. The 15 participants retained for the low memory capacity group had obtained a total score on the digit span test of 14 or less. To reduce the problem of the definition of the memory capacity being dependent on the sample of participants, the same criteria defining the memory groups were kept for Experiment 5 (people who obtained scores greater than 20 or lower than 14). This ensures a more stable definition of what is considered to be high and low individual memory capacity.

Method

Participants. Only participants who obtained a total score on the digit span test higher than 20 or lower than 14 were included in the analyses. They were 30 volunteer students (27 females and 3 males) from the University of Ottawa (29 undergraduates and 1 graduate student), between 19 and 42 years of age ($\overline{X} = 24$).

Apparatus, stimuli and procedure. Experiment 4 reprised Experiment 3 exactly, where participants were asked to judge of the efficacy of one, two or four medications presented simultaneously on a disease inducing skin discoloration. However, in Experiment 4, prior to the start of the experimental task (as described in Experiment 3), participants' memory capacity was assessed using a standard digit-span test (forward and backward). Fifteen participants who obtained a digit-span total score of 20 or higher were

selected for the high memory capacity group and fifteen participants who scored lower than 14 were selected for the low memory capacity group.

Results and Discussion

A *t* test analysis was first used to confirm that the total scores on the digit-span test obtained by the two memory capacity groups were significantly different from one another. The t test analysis confirmed that the scores of participants selected for the high memory capacity group ($\overline{X}$ = 22.67, *SD* = 1.76) were significantly higher than the scores obtained by the low memory capacity group ($\overline{X}$ = 12.80, *SD* = 1.21), $t(28) = -17.91$, $p = .00$.

All data were analysed using repeated-measures analyses of variance (ANOVA) with a Type I error rate of .05. To protect against possible violations of sphericity assumptions, a Huynh–Feldt correction was applied to all repeated measures with >1 df in the numerator. Because there is no clearly identified and statistically valid error term with mixed designs permitting to calculate post-hoc tests, such analyses were not included in this plan.

The main results for the contingency judgements of target contingencies are depicted in Figure 12. The slopes of the lines in the graph seem to indicate that contingencies of .40 and .60 were well discriminated across conditions. Also, while it seems that the number of simultaneous contingencies had an impact in the low memory capacity group with the presence of two or four simultaneous contingencies decreasing judgements (reproducing the effect obtained in Experiment 3), it looks as though the effect is absent in the high memory capacity group where there appear to be no difference

between judgements of target contingencies in cases where they were presented alone, with one additional contingency or with three additional contingencies.

Insert Figure 12 approximately here

Statistical analyses corroborate these impressions. A factorial analysis of variance with one between-subjects factor (2 groups of memory capacity) and two within-subjects factors (3 number of simultaneous contingencies by 2 values of target contingencies) revealed a marginally significant main effect of memory capacity group (F (1, 28) = 4.06, MSE = 2354.93, p = .054, η^2 = 0.13), a main effect of the number of simultaneous contingencies (F (1.76, 49.16) = 18.03, MSE = 737.06, η^2 = 0.39), a main effect of target contingencies (F (1, 28) = 9.04, MSE = 854.33, η^2 = 0.24) and an interaction of memory capacity group by number of simultaneous contingencies (F (1.76, 49.16) = 737.06, MSE = 647.08, η^2 = 0.15). No other results were significant. The main effect of target contingencies show that contingencies of .40 ($\overline{X}$ = 36.46) and .60 ($\overline{X}$ = 49.55) were well discriminated across all conditions. The main effect of memory capacity group show that judgements of target contingencies were lower in the low memory capacity group ($\overline{X}$ = 35.72) as compared with the high memory capacity group ($\overline{X}$ = 50.29). The main effect of simultaneous contingencies also confirms that judgements of the target contingencies were higher when participants assessed one contingency rather than two or four simultaneous contingencies. These results replicate those obtained in Experiment 3 and support Wagner's hypothesis regarding memory limitations to the activation of potentially up to three nodes fully in their *A1* state.

The interaction of memory capacity group and number of simultaneous contingency show that judgement of the target contingencies were underestimated when judging two or four simultaneous contingencies (comparatively to one) but mostly for the low memory capacity group. As stated previously, it is very well possible that people with a higher memory capacity also have a higher capacity for maintaining more elements simultaneously in their primary state of activity (*A1*). In other words, it is possible that they possess a higher *C1* constant. Consequently as the number of elements simultaneously in *A1* state rises, there would be less need for the probability of decay to increase. As a result, associative strength would be free to build according to its normal rate, even when the number of nodes activated is large. This is reflected by the greater deterioration in judgements of the target contingencies observed in the low memory capacity group.

A second set of analyses was conducted on the judgements of the target contingencies when they were presented with three other simultaneous contingencies in order to study the impact of the value of the additional contingencies. Results are shown in Figure 13. The positions of the curves seem to indicate that contingencies of .40 and .60 were well discriminated. The lower curves in the panel for the low memory capacity group seem to indicate that judgements for the target contingencies were lower in the low memory capacity group (compared to the high memory capacity group). Finally, especially in the panel representing the results for the high memory capacity group, it seems that judgements for the target contingencies increased as the values of the additional contingencies increased. A factorial analysis of variance with one between-subjects factor (2 groups of memory capacity) and two within-subjects factors (4 values

of additional contingencies by 2 values of target contingencies) confirm these impressions. It revealed a main effect of the memory capacity group (F (1, 28) = 10.32, MSE = 3256.49, η^2 = 0.27), a main effect of the target contingency (F (1, 28) = 7.40, MSE = 1540.05, η^2 = 0.21) and an interaction of memory capacity group by value of additional contingencies (F (2.82, 78.93) = 3.99, MSE = 1446.71, η^2 = 0.12). No other effect was significant. The main effect of memory capacity group shows, once again, that contingency judgements given by participants in the low memory capacity group were significantly lower ($\overline{X}$ = 19.95) than judgements made by the high memory capacity group ($\overline{X}$ = 43.62). This is an indication of the potential faster decay of elements in the low memory capacity group, which would impede the formation of associative strength. The main effect of the target contingency shows that participants discriminated between the contingencies of .40 ($\overline{X}$ = 24.89) and .60 ($\overline{X}$ = 38.67). The significant interaction reveals that the pattern in the two memory capacity groups regarding the impact of the values of the additional contingencies on the judgement of the target contingencies is different. While the target contingencies seem to have been judged lower when accompanied by lower values of additional contingencies (e.g., 0 and .20 versus .80 and 1), replicating earlier results, this contextual effect only appears to be present in the high memory capacity group. A different pattern is observed in the low memory capacity group, where judgements seem to show the reversed contextual effect over the 0/.20, .20/.80 condition, with a deflection in the .80/1 condition. This effect will be discussed in the General Discussion. Because Experiment 4 essentially repeated Experiment 3, it was once again impossible to investigate the effect of adjacent contingencies specifically, as position of contingencies was free to vary and different values of additional contingencies

were used. However, the significant results obtained in this experiment, as well as in the previous one, show that the contextual effects may not be limited to adjacent contingencies, but rather, all contingencies presented within a trial.

Insert Figure 13 approximately here

Confidence estimates were also analysed. The main results are shown in Figure 14. Inspection of the graph seems to reveal a decrease in confidence as the number of simultaneous contingencies increased. This effect seems more pronounced in the low memory capacity group. Furthermore, the slopes of the lines seem to indicate that confidence increased with the value of contingency. Statistical analyses confirm this. A factorial analysis of variance with one between-subjects factor (2 groups of memory capacity) and two within-subjects factors (3 number of simultaneous contingencies by 2 values of target contingencies) revealed a marginally significant main effect of memory capacity group (F (1, 28) = 4.05, MSE = 1469.34, p = .054, η^2 = 0.13), a main effect of number of simultaneous contingencies (F (1.56, 43.60) = 31.76, MSE = 326.13, η^2 = 0.53) and a marginally significant main effect of target contingency (F (1, 28) = 4.19, MSE = 132.46, p = .05, η^2 = 0.13).

Insert Figure 14 approximately here

The main effect of memory capacity group indicate that participants in the low memory capacity group were less confident in their judgements of the target contingencies ($\overline{X}$ =

56.66) than those in the high memory capacity group ($\overline{X}$ = 68.16). The main effect of number of simultaneous contingencies show that, regardless of the memory capacity group to which participants belonged, confidence significantly decreased as number of contingencies increased. Therefore, although participants in the high memory capacity group showed more confidence, in general, than those in the low memory capacity group, they also showed a decrease in confidence as number of simultaneous contingencies increased. In addition, the main effect of target contingency show that participants were more confident in judging a contingency of .60 ($\overline{X}$ = 64.17) rather than .40 ($\overline{X}$ = 60.65). These results confirm previous ones where confidence increased with the ΔP value.

Once again, a second set of analyses were performed on the subset of judgements for the target contingencies when they were presented with three additional contingencies (for a total of four contingencies presented simultaneously). Statistical analyses revealed a main effect of additional contingency (F (3, 84) = 5.89, MSE = 477.74, η^2 = 0.17). No other results were significant. The results of these analyses are presented in Figure 15. An inspection of the graph seems to indicate that confidence was higher for judgements of target contingencies when these were accompanied with stronger contingencies (e.g., .80 or 1). Recall that, as observed in previous results, confidence is usually higher when the contingency is stronger. It is therefore possible that, since people are more confident when assessing stronger contingencies, their level of overall confidence in series of trials that contained stronger contingencies was also higher. As a result, this may have increased participants' level of confidence for the target contingencies when they were presented within such series.

Insert Figure 15 approximately here

Once again, the results obtained in Experiment 4 support Wagner's hypothesis regarding the influence of memory capacity on contingency judgements. As the number of contingencies – and hence the number of nodes simultaneously activated in memory – increased, contingency judgements decreased, as predicted by the theory. As explained previously, this is due to the fact that, when there are too many elements simultaneously in their primary activity (or A1) state, this will be compensated for by a quicker decay of a portion of elements across all memory nodes. Consequently, associative strength cannot build as much as it normally would and contingencies are underestimated (as compared to presenting only one contingency). As the memory capacity, according to Wagner's theory, is taxed by increasing the number of simultaneous contingencies, we should also expect to see an impact of multiple contingencies on individual memory capacity. This is indeed what was observed in Experiment 4. A greater deterioration in judgements was observed in the low memory capacity group (as compared with the high memory capacity group) as number of simultaneous contingencies increased. As expected, participants in the low memory capacity group were also less confident in their judgements (than those in the high memory capacity group) and confidence decreased with increasing number of simultaneous contingencies. Once again, we also observed that confidence increased with ΔP. This effect will be revisited in the General Discussion.

Experiment 5

According to Kareev's view, individual differences in memory capacity should have an impact on contingency judgements based on series containing an equal number of events. This would occur because variations in individual memory capacity would lead participants to remember different sizes of samples from the series of events. Recall that Kareev implies that people attempt to memorize specific events in order to assess contingencies. Since they cannot remember all events ever experienced, they will base their judgement on a sample of events. Any judgement based on a sample should overestimate the contingency as a sample's correlation tends to be larger than that of the population from which it was taken. The smaller an individual's memory capacity, the smaller the sample of events that is remembered and consequently, the higher should be its corresponding contingency judgement. Participants who possess a low memory capacity should therefore overestimate contingencies between stimuli.

On the other hand, In Wagner's view, individual memory capacity would have no impact on estimating contingencies based on series containing an equal number of events. As there is no need for the recall of episodic events, memory capacity would not be tested with this manipulation. Estimates should therefore be equal in both groups. However, in Wagner's perspective, individual memory capacity should have an impact when assessing *simultaneous contingencies* and this is what was observed in Experiment 4.

Although the evidence acquired thus far favours Wagner's model, the fact remains that Kareev et al. (1997) had obtained a memory effect in their study (Experiment 1) when participants assessed contingencies based on series of equal length. Recall that, the authors reported that the perception of correlation was "more extreme" among

participants with a low working memory capacity as compared with participants who possessed a high working memory capacity. However, these results remain suggestive as two major problems with this study were identified. First, participants' perception of correlation was not collected. Instead, participants were asked to predict the content (X or O) of coloured envelopes (red or green). A contingency coefficient (φ) was calculated using participants' predictions. This procedure is equivalent to computing contingencies based on participants' estimates of probability and, as demonstrated by Wasserman, Elek, Chatlosh, and Baker (1993), there is not a good match between probability estimates and ΔP (which is equivalent to φ) nor between probability estimates and direct contingency estimates. A second problem is that, because the researchers calculated a contingency coefficient (φ) using participants' probability estimates, the bias obtained may reflect a bias that is part of this calculated correlation but that may not be part of the subjective perception at all (if we asked participants to evaluate the contingency directly). Subjective correlation is likely not computed cognitively using a statistical coefficient such as φ or ΔP but rather might coincidentally result from an analogous process based on associative strength. In light of these issues, it is not clear if the results obtained by Kareev et al. (1997) portray a real effect or are just an artefact resulting from the calculation of a contingency coefficient. Even if the obtained effect is genuine, it might actually have been caused by a need to store two contingencies in memory rather than by the sample size per se. Since participants were asked to predict the envelope content (X or O) on the basis of the envelope colour (red or green), in effect they were required to track the frequencies of two contingency tables, that of the actual envelopes and that of their own predictions in order to adjust their prediction performance. Experiment 5

investigates Kareev's predictions while correcting the problems of the Kareev et al.'s (1997) study by asking participants to assess contingency directly, one contingency at a time. Negative contingencies are also tested for generality.

Method

Participants. Participants were 30 volunteer undergraduate students (22 females and 7 men) from the University of Ottawa, aged between 18 and 38 ($\overline{X} = 24$).

Apparatus. The experimental task was administered using an IBM compatible microcomputer, equipped with a 14 inch VGA monitor, which was also used to collect the data. Programming of the study was done using Microsoft Visual Basic Professional 6.0.

Stimuli. The same stimuli used in previous experiments were also used in Experiment 5. Participants were asked to assess the effect of different medications on a disease inducing skin discoloration. However, contrary to Experiments 2 and 3, they only assessed a single contingency at a time. The medical files remained visible for 1 s with 1 s inter-trial intervals. Half of the participants were shown a red pill and the other half was shown a blue pill. In each case, the skin discoloration matched the pill's colour.

Procedure. Upon their arrival at the laboratory, participants were led to an individual testing room where they were informed of the experimental procedures on a computer screen. The experimenter read the following directions:

> Welcome to the Cognitive Psychology Laboratory. We are studying decision making processes such as those used by health professionals. Imagine that you have access to the medical files of many people sick with different diseases. These diseases have many symptoms, including one in common: a skin discoloration.

> Different medications treating these diseases have been used in clinical trials. The problem is that the medications themselves can also increase or decrease the risk of skin discoloration. We will ask you to evaluate different medications. For each medication, the clinical trial contains the results of many individual files. Each file may report absence of medication or presence of medication, the presence of skin discoloration or the absence of skin discoloration, in different possible combinations [the different combinations were shown to the participants]. You will have an opportunity to warm up your diagnostic skills with three practice trials.

These instructions were followed by three series of practice trials. A total of 20 medical files were shown for each contingency involved in the practice trials. The first practice has a contingency ΔP of .67. Using a 2 x 2 contingency table where cell *a* represents the frequency of the presence of both events, cell *b* represents the frequency of the presence of the treatment but the absence of skin discoloration, cell *c* represents the frequency of the absence of treatment but the presence of skin discoloration, and cell *d* the frequency of the absence of both events, the frequencies for the first series of practice trials were 10, 5, 0 and 5 for cells a, b, c, and d respectively. Trial order was randomized. Practice two used a contingency of -.67 with frequencies of 0, 5, 10, and 5. The third practice used a contingency of 0 with frequencies of 5, 5, 5, and 5.

At the end of each practice, participants were asked to judge the efficacy of the treatment relative to the absence of the treatment based on all files presented for the clinical trial. On the same screen, participants were also asked to assess their degree of confidence in their answer. The same judgement scales as before were used. The

experimenter was present for the practice trials in order to answer questions. No direction was given as to suggest any one particular way of arriving at an answer. Participants were told that they might have the impression of sometimes guessing the answer but that this impression was normal. It was also emphasized that the relations between treatments and effects were constant within series of clinical trials but varied between series. Data from the series of practice trials were omitted from the analyses.

After ensuring clear understanding of the task, the experimenter left the room and participants began the experimental series. A total of 8 series of clinical trials were presented randomly to the participants. Each series of trials contained 20 medical files. Series of 20 medical files were selected as opposed to 40 because, assuming that Kareev's view is accurate and people sample a series when its length is longer than their working memory capacity, 20 events would surely exceed participants' highest memory capacity and should lead to the sampling effect hypothesized by Kareev. Longer series are therefore not required. The test series used contingencies of .20, .40, .60, .80, -.20, -.40, -.60 and -.80. For the contingency of .20, the frequencies for cells *a*, *b*, *c*, and *d* were 6, 4, 4, 6, respectively. For the contingency of .40, frequencies were of 7, 3, 3, 7; for .60 they were 8, 2, 2, 8; for .80 they were 9, 1, 1, 9; for -.20 they were 4, 6, 6, 4; for -.40 they were 3, 7, 7, 3; for -60 they were 2, 8, 8, 2; and for -.80 they were 1, 9, 9, 1. At the end of each series of medical files, the judgement/confidence screen appeared with the rating scales. The contingency judgement scale was reset to 0 (middle of the scale) and the confidence scale was reset to 50 (middle of the scale) for each new judgement screen. Participants could change their answers before moving on to the next clinical trial but they could not return to their current judgement screen after submitting their responses.

Before each new series, a warning window appeared reminding participants to start their evaluation afresh as they were about to start a new series of clinical trials.

Prior to the start of the experiment, a standard digit-span test (forward and backward) was given to participants in order to measure their individual memory capacity. Fifteen participants who obtained a total score on the digit span test higher than 20 (as per the criteria established in Experiment 4) were selected for the high memory capacity group and 15 participants who obtained a total score on the digit span test lower than 14 were selected for the low memory capacity group. Participants who did not fall within these ranges were invited to participate in a different, but similar, experiment ongoing at the same time.

Results and Discussion

A *t* test analysis was first used to confirm that the total scores on the digit-span test obtained by the two memory capacity groups were significantly different from one another. The *t* test analysis confirmed that the scores of participants selected for the high memory capacity group ($\overline{X}$ = 21.80, *SD* = 2.54) were significantly higher than the scores obtained by the low memory capacity group ($\overline{X}$ = 13.00, *SD* = 1.56), $t(23.23) = -11.43$, $p = .00$.

All data were analysed using repeated-measures analyses of variance (ANOVA) with a Type I error rate of .05. To protect against possible violations of sphericity assumptions, a Huynh–Feldt correction was applied to all repeated measures with >1 df in the numerator. Because there is no clearly identified and statistically valid error term with mixed designs permitting to calculate post-hoc tests, such analyses were not included in this plan.

The main results for the contingency judgements of target contingencies are depicted in Figure 16. The contingency estimates for the negative contingencies were multiplied by -1 for the purpose of analyses, to render them comparable to the judgements for the positive contingencies. The slopes of the lines in the graph seem to indicate that, in general, the contingencies were well discriminated. It also looks as though some contingencies were underestimated in the low memory capacity group (in comparison to judgements in the high memory capacity group), and especially when contingencies were strong. Statistical analyses confirm these impressions. A factorial analysis of variance with one between-subjects factor (2 groups of memory capacity) and two within-subjects factors (2 valences: positive and negative contingencies; by 4 values of contingencies: .20, .40, .60 and .80) revealed a main effect of value of contingency (F (3, 84) = 16.19, MSE = 1607.50, η^2 = 0.37) and an interaction of memory capacity group by values of contingency (F (3, 84) = 3.75, MSE = 1607.50, η^2 = 0.12).

The memory capacity by value of contingency interaction effect support the visual impression that contingencies were sometimes underestimated by the participants in the low memory capacity group. This effect seems to be stronger with stronger contingencies. Kareev predicted that the low memory capacity group would *overestimate* contingencies as the samples remembered would be smaller. The opposite effect was found.

Insert Figure 16 approximately here

Confidence estimates were also analysed. The results are depicted in Figure 17. The slopes of the lines seem to indicate that confidence, once again, increased as ΔP value increased. Analyses confirm this: they revealed a main effect of value of contingency (F (3, 84) = 19.01, MSE = 162.79, η^2 = 0.40). No other effect was significant. These effects robustly replicate those obtained in the previous experiments.

Insert Figure 17 approximately here

Results of Experiment 5 contradict, once again, Kareev's predictions. Kareev predicted that the low memory capacity group would overestimate contingencies (compared to the judgements of the high memory capacity group) due to the smaller samples of events that are remembered. Results show that the contingencies in the low memory capacity group were in fact *underestimated* compared to the judgements of the high memory capacity group, especially when the contingency was strong. Although these results clearly contradict Kareev's hypothesis, they can be readily explained by Wagner's theory. As mentioned previously, it is possible that the *C1* constant varies with memory capacity. The constant may well reflect people's memory capacity. If people in the low memory capacity have a lower *C1* constant, it is possible that their constant increases the probability of decay of some elements in *A1* state, even with a single contingency present. Because elements do not remain in their *A1* state as long as they

should, associative strength cannot build as it should either. Consequently, contingencies are underestimated, especially with shorter series of 20 trials as were used in the current experiment. However, although associative strength may accumulate more slowly in participants who possess a lower memory capacity, according to Wagner's theory, with enough trials, associative strength should still eventually reach asymptote. Returning to Figure 12, we can see that when participants were evaluating only one contingency, the positions of the lines in the graph do seem to be at the same height for the low and high memory capacity groups. This indicates that when a single contingency is presented, after viewing 40 trials, judgement in the lower memory capacity group has reached the same level as that in the high memory capacity group (which probably indicates that asymptote was then reached). Twenty trials would, however, be insufficient to permit judgements in the lower memory capacity group to reach asymptote.

General Discussion

The current thesis aimed at clarifying the role of memory and memory capacity as involved in contingency judgements. Kareev (1995, 1997) implicitly assumes that people attempt to memorize the sequential events that form a contingency (i.e., the effect with the presence or absence of the cue and the absence of the effect with the presence or absence of the cue) in order to arrive at a contingency judgement. According to this view, increasing the length of the sequential events contained in a contingency would therefore tax memory capacity. In this perspective, Kareev also assumes that, because memory capacity is limited, people will base their judgement on a *sample* of events (since, unless the series are short, people cannot possibly remember all the events leading to a contingency). Because the sampling distribution of the Pearson correlation coefficient

tends to be skewed when the correlation of the population is different from zero and the skewness increases with reductions in sample size, Kareev argues that the sampling of a reduced number of events, or a smaller working memory capacity, should lead people to *overestimate* the normative relations between events. As the detection of covariation is essential to people's adaptation to their environment, this mechanism would prove very helpful as, due to their overestimation, covariations should also be more easily detected. While Cheng's theory makes no predictions regarding memory as involved in contingency judgements, because of its statistical nature (i.e., calculation of a ΔP coefficient), Kareev's predictions are entirely compatible with it. The distribution of ΔP is also subject to the same bias found in the Pearson's correlation. However, Cheng's theory could also predict lower judgements with smaller samples (as compared with larger samples) due to reduced confidence in evaluating smaller samples.

Wagner's view is quite different. According to him, people do not attempt to memorize the sequence of events in a contingency, as they do not need to do so in order to arrive at a contingency judgement. Instead, Wagner proposes that people assess contingencies based on the associative strength that forms between the internal representations of two (or more) stimuli (represented in the form of memorial nodes). Wagner suggests that what stresses memory capacity is not an increase in the length of the series of episodes as proposed by Kareev (since, in Wagner's view, people do not need to memorize the episodes), but rather, an increase in the number of contingencies that are presented simultaneously to the cognitive system (as this will increase the number of memory nodes simultaneously active). Wagner postulates memory limitations in the form of the number of nodes that can be simultaneously active.

Wagner's (SOP) and Kareev's (narrow window approach) predictions were tested in five experiments. Experiment 1 tested Kareev's notion that smaller samples would lead to an overestimation of a covariation by manipulating sample size directly. Results contradicted Kareev's prediction: With smaller samples, participants *underestimated* the contingency rather than overestimating it (as compared with judgements of larger samples). Wagner's theory, on the other hand, can readily account for this result because a smaller number of trials should lead to less associative strength, resulting in an underestimation of contingencies. Although Cheng's theory could also account for these results by proposing that lower estimates with small samples are solely the product of combined contingency judgement and reduced confidence, this proposition was contradicted in Experiment 1. Indeed, while contingency judgements increased with increasing sample size, confidence decreased. While Kareev predicted that people with lower memory capacity would also overestimate contingencies (as their remembered sample would be smaller), this prediction was also contradicted in Experiment 5.

Experiment 2 and 3 tested Wagner's assumptions by increasing the number of contingencies that were presented simultaneously. Wagner proposes that only two or three nodes could be fully in their A1 state simultaneously and that exceeding these limits would cause a portion of elements contained in each activated node to decay quicker. Following this assumption, when people have to assess more than one contingency (which itself activates a total of three nodes) at once, their judgements should be reduced (as a quicker decay of the elements contained in the nodes would prevent associative strength from building fully). This is what was observed. Wagner's predictions were also supported in Experiment 4, which compared judgements between different memory

capacity groups. The reduction in judgements caused by an increase in the number of contingencies presented simultaneously was greater for participants with a low memory capacity.

In addition to the main results of the current thesis supporting Wagner's SOP theory regarding the best way to tax memory capacity, other interesting results were found. First were the surprising results observed from people's contingency estimates: 1- the absence of further deterioration in judging more than two simultaneous contingencies; and 2- the contextual effects seen through the influence of additional contingencies on an evaluated contingency. Furthermore, interesting results were also obtained with confidence estimates: 1- the dissociation between contingency and confidence estimates; 2- the increase in confidence with ΔP value; and 3- the increase in confidence with sample size. These results are discussed in greater detail in the following sections.

Results Obtained with Contingency Estimates

While results generally supported Wagner's SOP theory, other results seem, for the time being, to fall outside the scope of its explanatory power. The absence of further decrease in contingency judgements beyond two simultaneous contingencies is a case in point. While Wagner's SOP proposes a mechanism by which memory load can be taxed and discusses the impact of this manipulation on contingency judgements, it does not consider other cognitive processes that might help people in estimating multiple contingencies. Indeed, in their natural environment, people are constantly facing and evaluating multiple simultaneous contingencies. Consequently, it would only be logical to expect that they would have evolved in developing a number of adaptive mechanisms to accomplish this task successfully. For example, as stated previously, it is possible that

people focus their attention on a single contingency at a time and consequently, base their estimate on a sample of events for each contingency. With a sufficient number of events, even judgements based on a sample could reach asymptote (or in the case of the results contained in this thesis, reach the level of performance found with two or more simultaneous contingencies).

The possibility of having associations form in *A2* state (as opposed to only in *A1* state) also offers an interesting extension to Wagner's SOP theory. Indeed, as stated previously, if associations can be formed in the *A2* state (although weaker than in their *A1* state as they are on their way back to their inactive state), this could potentially account for the absence of further deterioration in judgements. While an initial deterioration was shown when presenting more than one contingency (because the limit of 2 or 3 nodes simultaneously in their *A1* state was reached), according to Wagner, a total of 10 to 15 nodes could be simultaneously in their *A2* state. Using contingencies with a single cause and a single outcome, as well as adding a node for the context, we may need as many as 5 to 7 simultaneous contingencies before contingency judgements show any further deterioration. This hypothesis could be tested in future studies.

Until recently, Wagner's SOP theory was also unable to account for contextual effects, such as those observed in the presence of multiple contingencies. For example, the model was unable to account for the context effect that Pavlov (1927) called *external inhibition*. In it simplest form, external inhibition is obtained in animal conditioning studies when a stimulus A is first reinforced and then presented in a test trial either alone or along with another stimulus B that is associatively neutral (ΔP=0). While the Rescorla-Wagner model does postulate that the context will activate a separate node, it predicts

that the response to stimulus A should be the same in both cases (whether it is presented with stimulus B or not). Recall that associative strength is represented by the sum of the associative strength of all stimuli present on a trial. If stimulus B is neutral, then the associative strength of A presented alone or in combination with B should be the same and hence, response should be the same. However, this is not what is observed. Instead, the responding to stimulus A presented with B is *less* than the responding to A alone. These results are very similar to those obtained in the current thesis where judgements of a contingency accompanied by a weaker one were also lower. Similarly, judgements were higher when a contingency was presented along with stronger contingencies. The impact of context on learning is now well documented: It has been observed in different areas of study with humans including memory and word recognition, among others.

In order to account for such contextual effects, Wagner (2003) proposed his replaced elements conception. Using a standard elemental theory conception, Wagner's SOP theory views memory nodes (which are the internal representations of stimuli) as collections of elements. In the replaced elements conception, it is proposed that nodes contain elements that are *context independent* (these elements are activated whenever their stimulus is presented to the cognitive system) and some that are *context dependent* (these elements will be activated or not depending on the presence or absence of other stimuli). When a stimulus is presented in combination with another, a proportion of its elements (that would be activated if the stimulus was presented alone) will be replaced by a portion of the elements of the accompanying stimulus. These elements will contribute to the formation of associative strength. It is therefore quite easy to see how a proportion of elements from an accompanying weaker contingency B could reduce associative

strength of a given contingency A (since the elements contributed by stimulus B would be weaker and therefore generate less associative strength than if A was presented alone with all of its elements contributing to the formation of associative strength). And vice-versa, when a contingency is presented along with a stronger one then, the elements of the stronger contingency should cause the evaluation of the other contingency to be higher (since associative strength will be higher). This extension to the SOP theory provides an elegant way to account for most of the context effects observed in this thesis.

The results obtained in Experiment 4 are, however, more difficult to explain. While target contingencies seem to have been judged lower when accompanied by lower values of additional contingencies than when accompanied by higher values of additional contingencies (e.g., 0 and .20 versus .80 and 1), replicating earlier results regarding the contextual effects, the effect seems to only be present in the high memory capacity group. A different pattern emerges in the low memory capacity group, where judgements seem to show the reversed contextual effect over the 0/.20, .20/.80 condition, with an exception in the .80/1 condition, where judgements were higher. At this point, it is difficult to see how Wagner's replaced elements conception can account for this phenomenon.

While fewer elements can be simultaneously activated with lower memory capacity (meaning that perhaps, as a consequence, a smaller proportion of elements of the target contingency would be replaced by elements from the context), this mechanism still cannot account for the observed results. Even if a smaller proportion of elements in the evaluated associative strength comes from the context, in cases where this latter possesses a higher contingency, it should still increase the perceived value of the contingency being evaluated, only maybe less so than in the case where people have a

higher memory capacity (where they can activate more elements simultaneously and therefore the context can contribute a greater proportion of elements to the associative strength). Furthermore, the context should also somewhat decrease the value of the perceived contingency when the contingency of the context is lower than the contingency being evaluated. But this is not what is observed. Even if we hypothesize that perhaps the memory limitations of people with a lower memory capacity act to exclude contextual elements in the formation of associative strength altogether (in order to reduce the number of elements simultaneously active), this only seemingly leads to the expectation that judgements would therefore be unaffected by the context. But, once again, this is not what is observed. Contingency judgements are indeed influenced and in the opposite direction than would be expected. However, it is also possible that elements from the context do not contribute to the formation of the associative strength of the contingency being evaluated but that, because elements from the context do not contribute, a different mechanism emerges. It is possible that participants would then keep the context more separate and evaluate contingencies in comparison to one another. In this view, it is possible that contingencies of .40 and .60 accompanied by contingencies of 0 and .20 would appear stronger (in comparison to the lower ones with which they are shown). And the opposite would happen when a contingency of .80 is present; in that case, contingencies of .40 and .60 would seem weaker. However, this explanation is only partial as it still does not account for the reason why contingencies of .40 and .60, presented with contingencies of .80 and 1 would then seem stronger. This therefore remains a topic for further investigations.

Results Obtained with Confidence Estimates

The results obtained through the confidence estimates were also very interesting. Until now, it was difficult to determine whether decreases in contingency judgements were due to a decrease in associative strength, a decrease in confidence or a mix of both. In Experiment 1, we observed that while contingency judgements increased with sample size, confidence estimates *decreased.* Furthermore, dissociation between contingency judgements and confidence was also observed in Experiment 3 where changes in contingency judgements and confidence did not always follow the same pattern. These results were also replicated in Experiment 4 and 5. Therefore, we now know that confidence and contingency judgements are indeed two separable psychological entities that do not necessarily covary with one another. This causes a problem for Cheng's theory as it is now unable to rely on confidence to account for its inability to predict the learning curve.

Also of interest, was the consistent pattern of results observed between ΔP and confidence: as ΔP value increased, confidence increased as well. A possible explanation of this phenomenon is obtained through a closer inspection of the chosen tables of contingencies throughout the experiments. Indeed, a look at the tables that were used to build the different ΔPs reveal that, for ΔP closer to zero, frequencies in the four cells of the contingency table were more equally distributed. For example, for a ΔP of zero, the frequencies used were 10, 10, 10, 10 for each of the *a*, *b*, *c*, and *d* cells in the 2 x 2 contingency table. As contingency increased, the frequencies of the chosen tables tended to be more concentrated in two cells. For example, for the contingency of 1, a table with frequencies of 20, 0, 0, 20 was used. As was said previously, it is therefore possible that

confidence was influenced by the diversity or complexity of the objective experience when ΔP was small. Since it is possible to build tables of different complexity for a constant ΔP [for example, for a ΔP of 0, we could have frequencies of 5-5-5-5 (more complex table), as well as frequencies of 0-10-0-10 and 10-0-10-0 (for tables of lower complexity)], a future study could test this concept further and verify if this is indeed what affected confidence estimates in the experiments contained in this thesis. The relation between complexity and confidence could clarify the nature of one determinant of confidence ratings in contingency judgement tasks.

One might also wonder what some of the other processes underlying the concept of confidence are. As previously stated, it is very possible that confidence is influenced by people's beliefs regarding their own abilities, regardless if these beliefs are accurate or not. For example, the phenomenon of increased confidence with smaller samples, as observed in Experiment 1, may well be explained by the fact that people may believe that they need to memorize the sequence of events contained in a contingency judgement in order to perform well in the task. Consequently, when shown fewer events (e.g., 4 events), they feel more confident that they can handle the task, as compared with longer series of events (e.g., 40 events), unaware that this task does not require memorization of the events at all. In light of the results of Experiment 1 showing that contingency estimates for smaller samples are actually underestimated (in comparison with larger samples), we could conclude that participants' performance is actually more accurate with larger samples (since associative strength has reached asymptote). However, people are less confident in their performance with larger samples, showing that people might not always have an accurate idea of their own processes or abilities. These suggestions

remain hypothetical for the moment. The concept of confidence will need to be further examined in future studies in order to improve understanding of the various mechanisms that may underlie it.

Implications and Future Directions

Wagner's model has proven very successful in explaining a number of phenomena. For example, evidence has shown that its proposition of prevention of stimulus rehearsal due to priming is accurate in accounting for phenomena like the short-term refractory-like effect in studies of habituation (Davis, 1969; Whitlow, 1975), the conditioned diminution of the unconditioned response in Pavlovian conditioning (Kimble & Ost, 1961; Kimmel, 1966), the less persistent memory of signalled versus unsignaled samples in a short-term memory paradigm (Terry & Wagner, 1975), and the apparent decrease in the associative learning occasioned by a signalled unconditioned stimulus in studies of blocking in Pavlovian conditioning (Kamin, 1969). Its hypothesis of gradual growth in associative strength has also been confirmed in studies with a reduced number of trials (e.g., see Experiment 1) and it has also been very successful in accounting for many variables that affect compound conditioning effects such as blocking and overshadowing (Siegel & Allan, 1996). In addition, its predictions regarding the memory capacity as involved in contingency judgements (the number of nodes simultaneously in their *A1* states) have been supported by results contained in this thesis.

Another strength of Wagner's approach is that the formulas included in his model do not run into problems experienced by formal mathematics that would be encountered by statistical theories who assume that people calculate a ΔP or an equivalent. The division by zero that sometimes occur when judging contingencies based on small

samples of events represents such a case. Since people can indeed arrive at a contingency judgement, even in those cases, it is becoming even clearer than before that people may indeed not calculate a statistical coefficient per se. Consequently, statistical models may eventually need to revise their approach.

While there may still exist situations for which the SOP theory offers no prediction, the extensions that have been proposed in order to account for 1- an absence of further deterioration beyond two contingencies; and 2- contextual effects; offer interesting and promising avenues of research. In addition, Wagner's model could also be further expanded to include other cognitive attributes such as attentional processes, that may perform a preliminary selection on the material to be assessed. Although the current thesis focused mostly on the use of positive contingencies, future studies should investigate the memory effects using negative contingencies as well. The role of confidence in contingency judgements, its determinants as well as a more formal definition of the concept could also represent areas of investigation for future studies. A greater detailing of the concept of confidence may permit researchers to see what its implications for Wagner's model would be. Finally, since Wagner's SOP theory does not currently include a mechanism to account for causal attributions, a further expansion of the theory could be the addition of an element of causality, perhaps through a reconciliation of Cheng's causal power concept integrated with SOP.

However, even if future research should lead us to conclude that SOP is not the proper way in which to view memory as involved in contingency judgements, the shear volume of research that the theory has generated and keeps generating will definitely

have contributed in increasing our knowledge regarding how the mind works. This, in itself, already makes it a successful theory.

References

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

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

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

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

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

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

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

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2). New York: Academic Press.

Baddeley, A. D. (1976). *The psychology of memory.* New York: Basic Books.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14,* 575-589.

Baker, A. G., Berbier, 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* (4th ed., pp. 85-116). Hillsdale, NJ: Lawrence Earlbaum Associates.

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

Beach, L. R., & Scopp, T. S. (1966). Inferences about correlations. *Psychonomic Science, 6,* 253-254.

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

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psycholgy: General, 121,* 177-194.

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

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

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

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

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

Cheng, P. W., Park, J., Yarlas, A. S., & Holyoak, K. J. (1996). A causal-power theory of focal sets. *The Psychology of Learning and Motivation, 34,* 313-355.

Crocker, J. (1981). Judgement of covariation by social perceivers. *Psychological Bulletin, 90,* 272-292.

Darcheville, J., Madelain, L., Buquet, C., Charlier, J., & Miossec, Y. (1999). Operant conditioning in the visual smooth pursuit in young infants, *Behavioural Processes, 46,* 131-139.

Davis, M. (1969). Effects of interstimulus interval length and variability on startle response habituation. *Journal of Comparative and Physiological Psychology, 72,* 177-192.

DeCasper, A. J., & Fifer W. P. (1980). Of human bonding: Newborns prefer their mother's voices, *Science*, 1174-1176.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology, 49B*, 60-80.

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

Estes, W. K. (1955a). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369-377.

Estes, W. K. (1955b). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62,* 145-154.

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

Hamilton, D. L. (1976). Cognitive biases in the perception of social groups. In J. S. Carroll & J. W. Payne (Eds.), Cognition and social behaviour. Hillsdale, NJ: Erlbaum.

Hays, W. L. (1963). *Statistics for psychologists.* New York: Holt Rinehart & Winston.

Jennings, D. L., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgements. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under Uncertainty: Heuristics and Biases* (pp. 211-230). Cambridge: Cambridge University Press.

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

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

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56,* 263-269.

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

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126,* 278-287.

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

Kelley, H. H. (1971). *Attribution in social interaction.* Morristown, N.J.: General Learning Press.

Kimble, G. A., & Ost, J. W. P. (1961). A conditioned inhibitory process in eyelid conditioning. *Journal of Experimental Psychology, 61,* 150-156.

Kimmel, H. D. (1966). Inhibition of the unconditioned response in classical conditioning. *Psychological Review, 73,* 232-240.

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

Marr, D. (1982). *Vision.* New York: Freeman.

Mazur, J. E., & Wagner, A. R. (1982). An episodic model of associative learning. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analyses of behaviour: Acquisition.* (pp. 3-40). Cambridge, MA: Ballinger.

McGeoch, J. A. (1942). *The psychology of human learning.* New York: Longsmans, Green.

Mercier, P., & Parr, W. (1996). Inter-trial interval, stimulus duration and number of trials in contingency judgements. *The British Psychological Society, 87,* 549-566.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

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

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language, *Infant Behavior and Development, 16,* 496-500.

Parr, W. V., & Mercier, P. (1998). Adult age differences in on-line contingency judgements. *Canadian Journal of Experimental Psychology, 52,* 147-158.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94,* 61-75.

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

Robinson, G. H. (1964). Continuous estimation of a time-varying probability. *Ergonomics, 7,* 7-21.

Schull, J. (1979). A conditioned opponent theory of Pavlovian conditioning and habituation. In G. H. Bower (Eds.), *The Psychology of Learning and Motivation* (Vol. 13). New York: Academic Press.

Schustack, M. W. (1988). Thinking about causality, In R. J. Sternberg & E. E. Smith (Eds.), *The Psychology of Human Thought* (pp. 92-115). New York: Appleton-Century-Crofts.

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

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

Shaklee, H., & Goldston, D. (1989). Development in causal reasoning: Information sampling and judgement rule. *Cognitive Development, 4,* 269-281.

Shaklee, H., & Mims, M. (1986). Development of rule use in judgements of covariation. In H. R. Arkes & K. R. Hammond (Eds.), *Judgement and decision making: An interdisciplinary reader* (pp. 495-506). Cambridge: Cambridge University Press.

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

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

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

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgement. In G. H. Bower (Eds.), *The psychology of learning and motivation* (pp. 229-261). London: Academic Press.

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

Shanks, D. R., Holyoak, K. J., & Medin, D. L. (1996). *The psychology of learning and motivation: Causal learning* (Vol. 34). London: Academic Press.

Shuford, E. H. (1961). Percentage estimation of proportion as a function of element type, exposure time and task. *Journal of Experimental Psychology, 61*, 430-436.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review, 3*, 314-321.

Solomon, R. L., & Corbit, J. D. (1974). An opponent-process theory of motivation. *Psychological Review, 81*, 119-145.

Terry, W. S., & Wagner, A. R. (1975). Short-term memory for "surprising" versus "expected" unconditioned stimuli in Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 1,* 122-133.

Towse, J. N., Hitch, G. J., & Hutton, U. (2000). On the interpretation of working memory span in adults. *Memory & Cognition, 28,* 341-348.

Trolier, T. K., & Hamilton, D. L. (1986). Variables influencing judgements of correlational relations. *Journal of Personality and Social Psychology, 50,* 879-888.

Wagner, A. R. (1976). Priming in STM: An information processing mechanism for self-generated or retrieval-generated depression in performance. In T. J. Tighe & R. N. Leaton (Eds.), *Habituation: Perspectives from Child Development, Animal Behaviour, and Neurophysiology*. Hillsdale, NJ: Lawrence Erlbaum.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behaviour. In N. E. Spear & R. R. Miller (Eds.), *Information Processing in Animals: Memory Mechanisms* (pp. 5-47). Hillsdale, NJ: Erlbaum.

Wagner, A. R. (2003). Context-sensitive elemental theory. *The Quarterly Journal of Experimental Psychology, 56B,* 7-29.

Wagner, A. R., & Brandon, S. E. (1989). Evolution of a structured connectionist model of Pavlovian conditioning. In S. B.Klein & R. R. Mowrer (Eds.), *Contemporary Learning Theories: Pavlovian Conditioning and the Status of Traditional Learning Theory* (6th ed., pp. 149-189). Hillsdale, NJ: Lawrence Earlbaum Associates.

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In M. S. Halliday & R. A. Boakes (Eds.), *Inhibition and learning* (pp. 301-336). San Diego, CA: Academic Press.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgements of contingency. *Canadian Journal of Psychology / Revue Canadienne de Psychologie, 19,* 231-241.

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

Whitlow, J. W. (1975). Short-term memory in habituation and dishabituation. *Journal of Experimental Psychology: Animal Behavior Processes, 1,* 189-206.

Table 1

Design of Experiment 2

	Target contingency	
Number of simultaneous contingencies	.20	.80
1	.20	.80
2	.20, .20 .20, .80	.80, .20 .80, .80
3	.20, .20, .20 .20, .20, .80 .20, .80, .80	.80, .20, .20 .80, .20, .80 .80, .80, .80

Table 2

Design of Experiment 3

Number of simultaneous contingencies	Target contingency	
	.40	.60
1	.40	.60
2	.20, .40 .40, .60	.40, .60 .60, .80
4	0, .20, .40, .60 .20, .40, .60, .80	.20, .40, .60, .80 .40, .60, .80, 1

Figure Captions

Figure 1. Combinations of the presence or the absence of a cause and an effect, as found in a 2 x 2 contingency table.

Experiment 1

Figure 2. Mean contingency estimates as a function of sample size.

Figure 3. Mean confidence as a function of sample size.

Experiment 2

Figure 4. Example of a clinical trial inside a series of three contingencies.

Figure 5. Mean contingency estimates of the target contingencies as a function of the number of simultaneous contingencies.

Figure 6. Mean contingency estimates as a function of the type of stimulus.

Figure 7. Mean contingency estimates as a function of the values of the target or additional contingencies.

Experiment 3

Figure 8. Mean contingency estimates of the target contingencies as a function of the number of simultaneous contingencies.

Figure 9. Mean contingency estimates of the target contingencies as a function of the values of additional contingencies. Because of the factorial design and the combinatorics of adding three supplemental contingencies to either a .40 or .60 target contingency with two supplemental values below or above the target, the combination .20 .40 .60 .80 occurs twice in the design (see Table 2).

Figure 10. Mean confidence in estimates of target contingencies as a function of the number of simultaneous contingencies.

Figure 11. Mean confidence in estimates of target contingencies as a function of the values of additional contingencies. Because of the factorial design and the combinatorics of adding three supplemental contingencies to either a .40 or .60 target contingency with two supplemental values below or above the target, the combination .20 .40 .60 .80 occurs twice in the design (see Table 2).

Experiment 4

Figure 12. Mean contingency estimates of the target contingencies as a function of memory capacity group and number of simultaneous contingencies.

Figure 13. Mean contingency estimates of the target contingencies as a function of memory capacity group and values of additional contingencies. Because of the factorial design and the combinatorics of adding three supplemental contingencies to either a .40 or .60 target contingency with two supplemental values below or above the target, the combination .20 .40 .60 .80 occurs twice in the design (see Table 2).

Figure 14. Mean confidence in estimates of target contingencies as a function of memory capacity group and number of simultaneous contingencies.

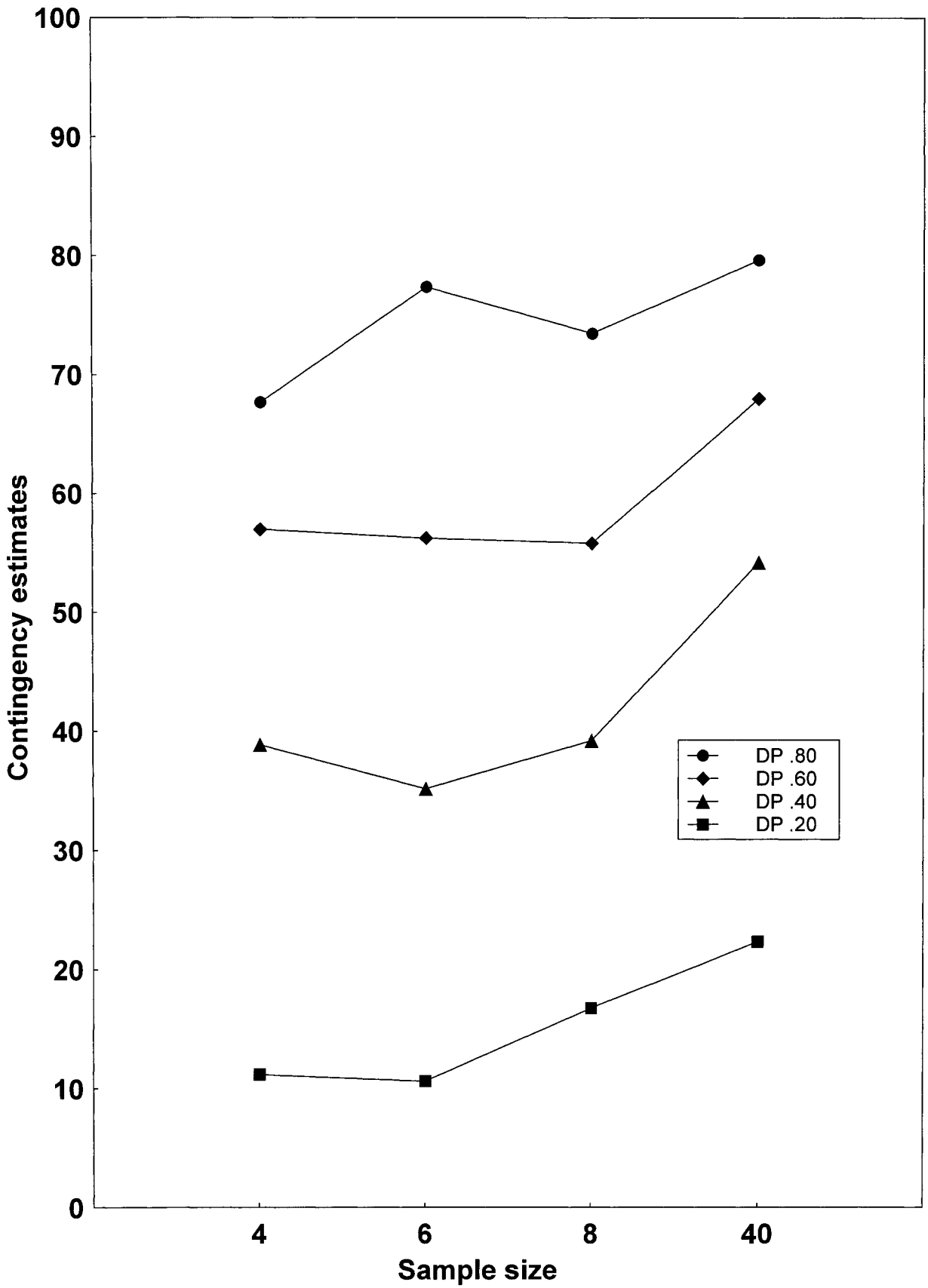
Figure 15. Mean confidence in estimates of target contingencies as a function of memory capacity group and values of additional contingencies. Because of the factorial design and the combinatorics of adding three supplemental contingencies to either a .40 or .60 target contingency with two supplemental values below or above the target, the combination .20 .40 .60 .80 occurs twice in the design (see Table 2).

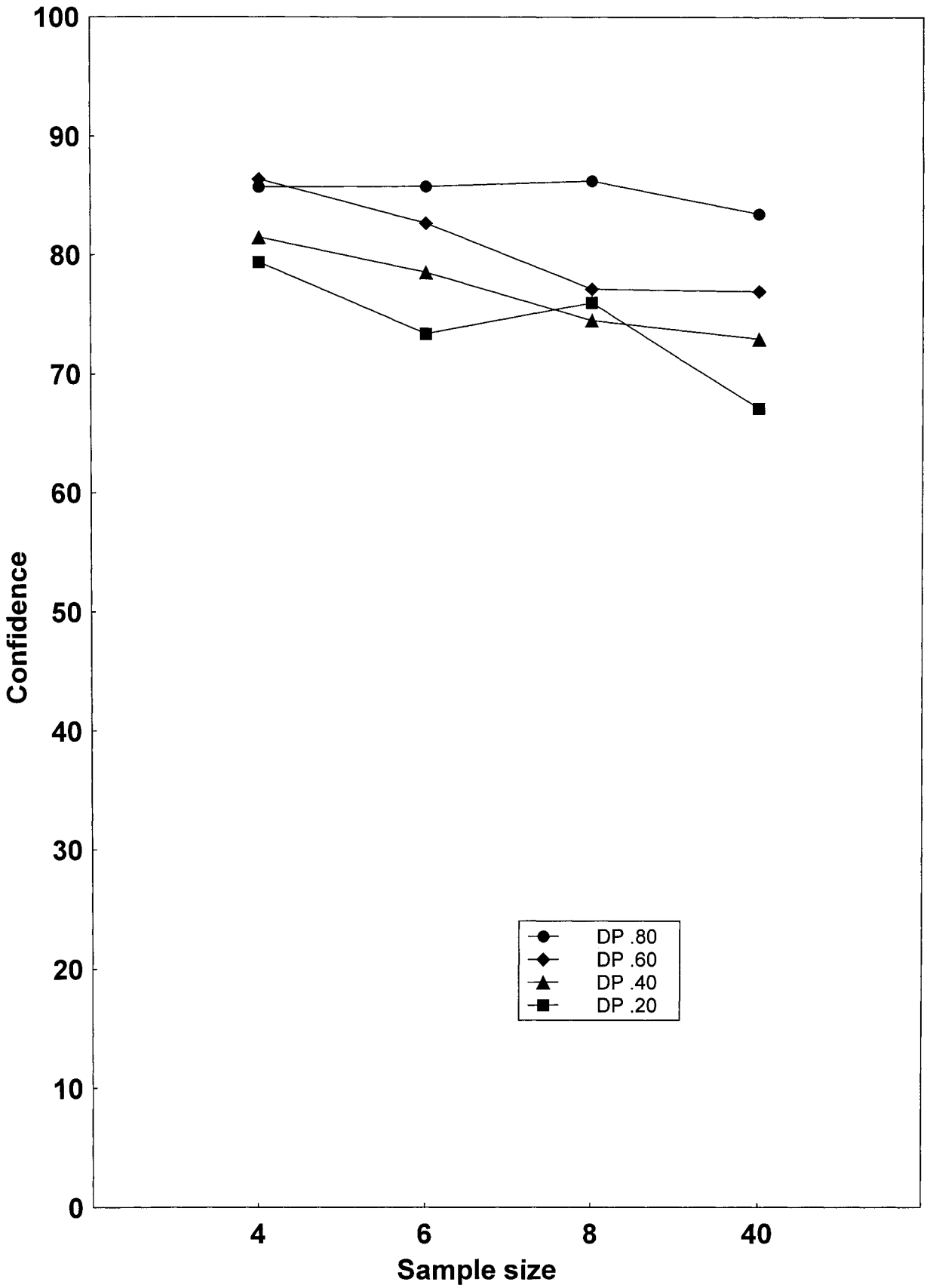
Experiment 5

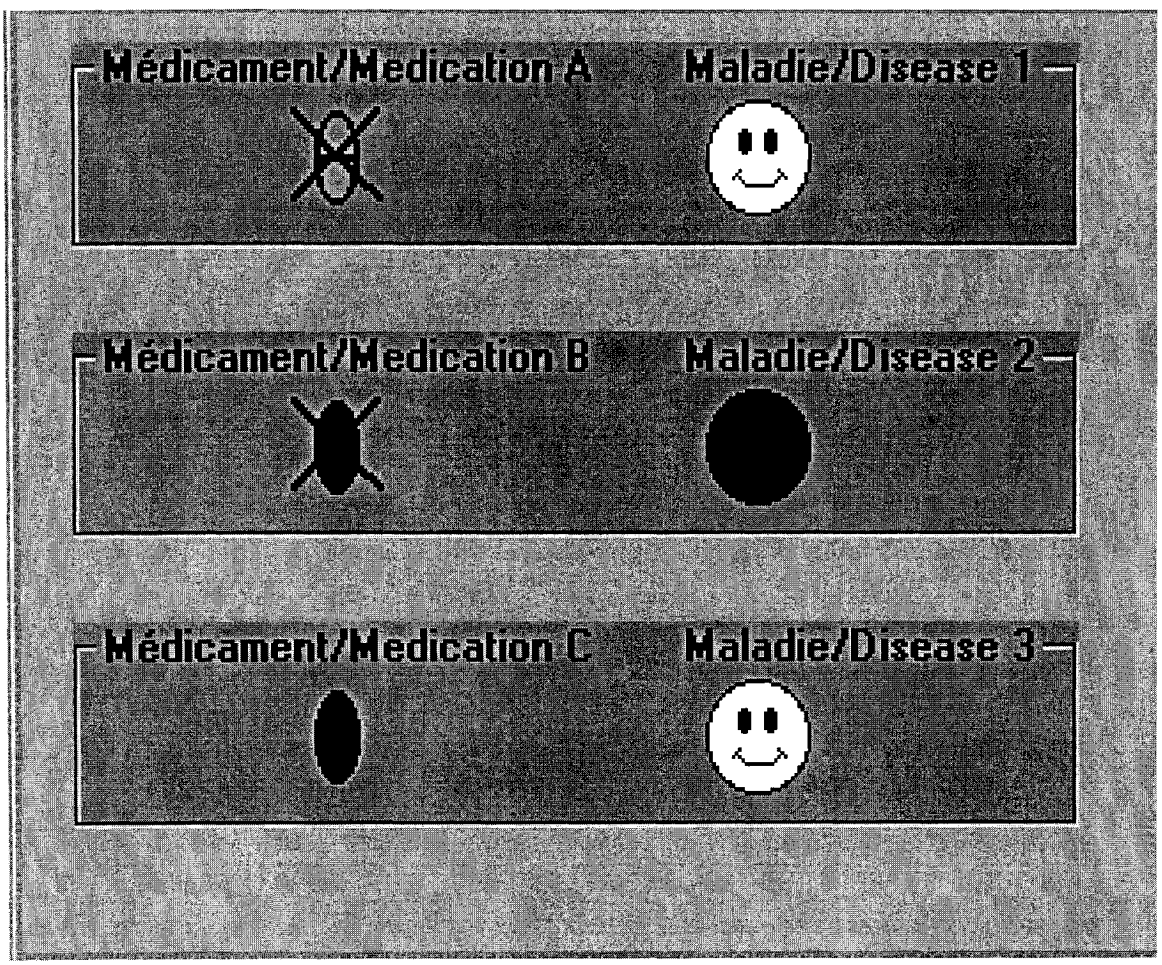
Figure 16. Mean contingency estimates of the contingency values as a function of memory capacity group and valence.

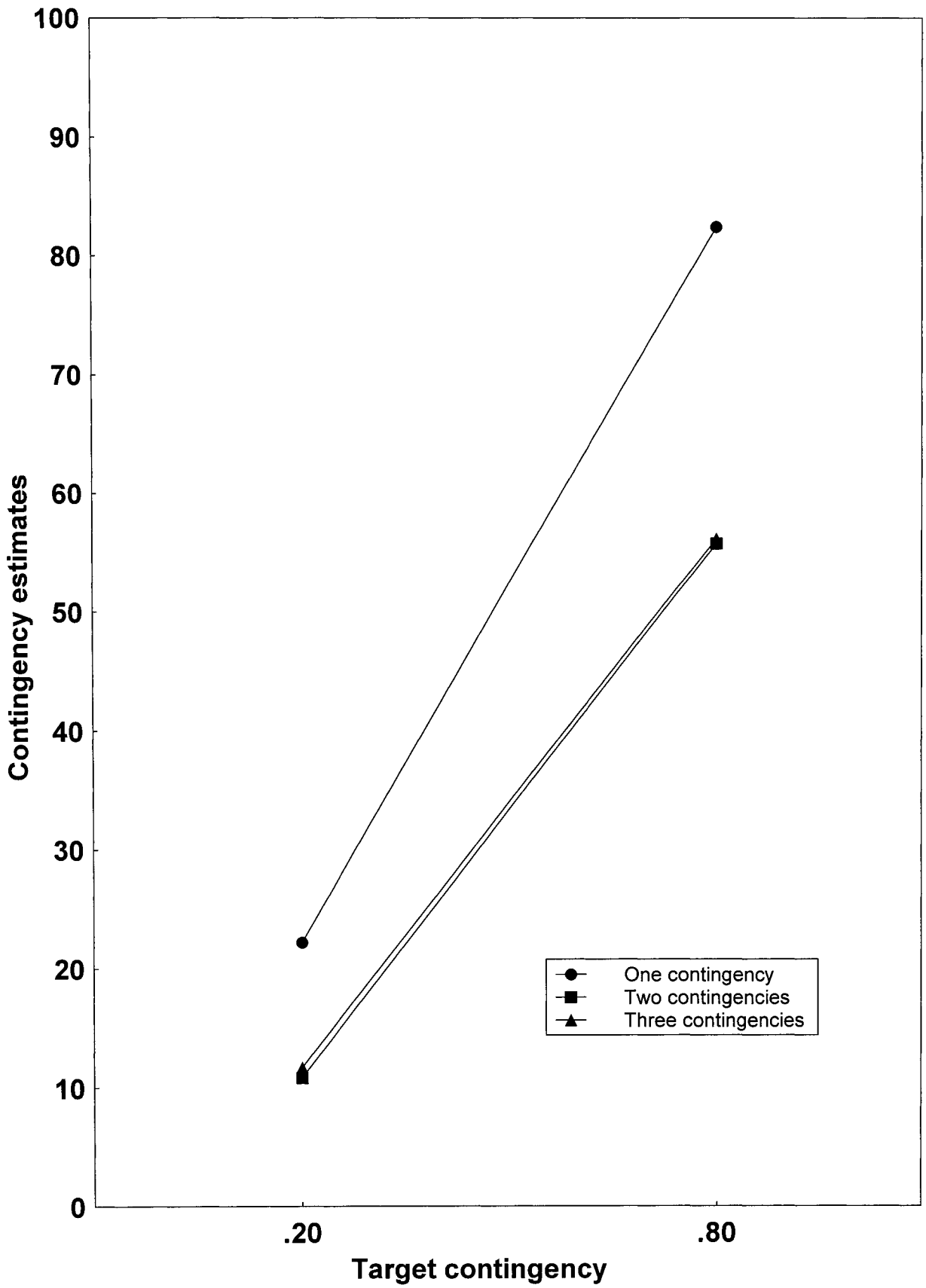
Figure 17. Mean confidence in estimates of the contingency values as a function of memory capacity group and valence.

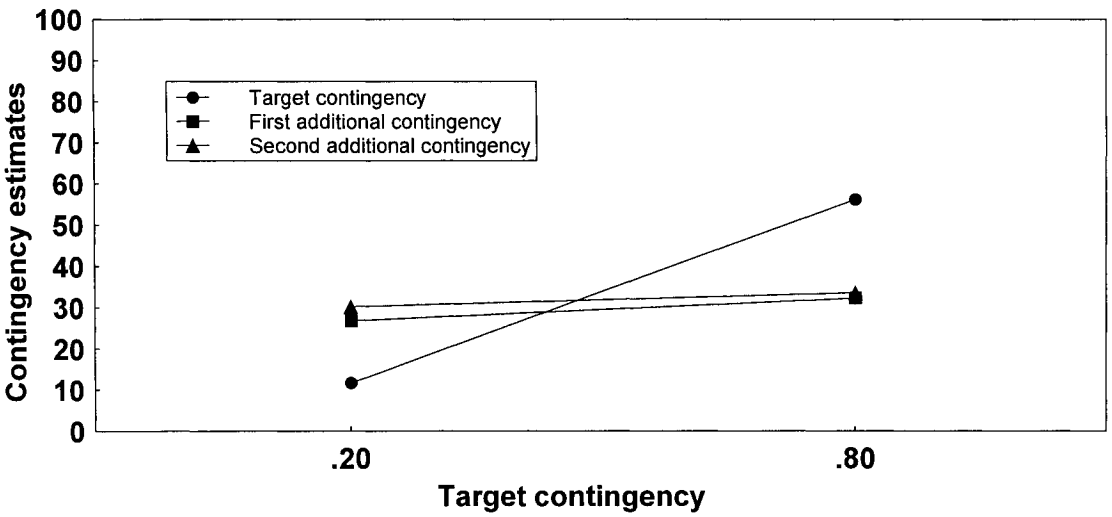
	Effect present	Effect absent
Cause present	A	B
Cause absent	C	D

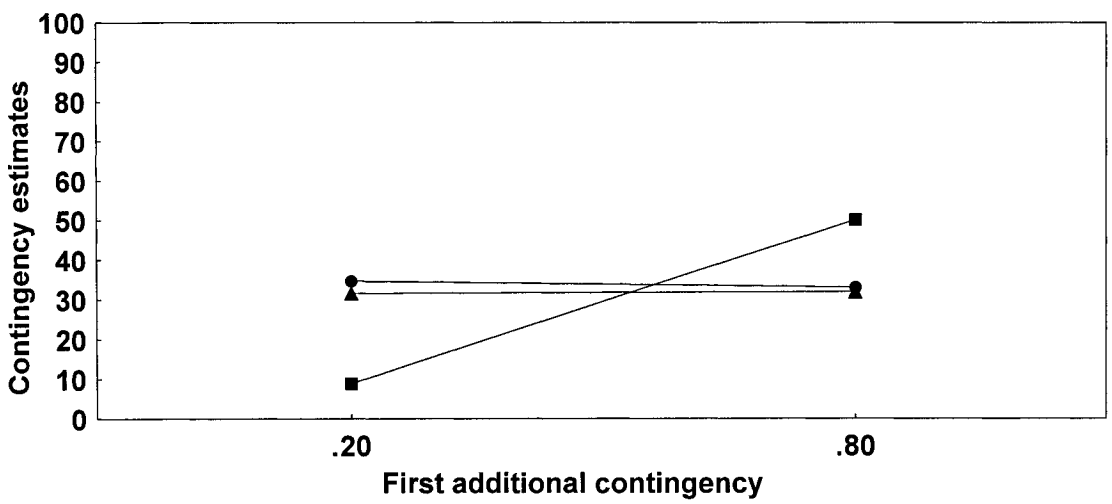

100
90
80
70
60
50
40
30
20
10
0
Contingency estimates
4
6
8
40
Sample size
DP .80
DP .60
DP .40
DP .20

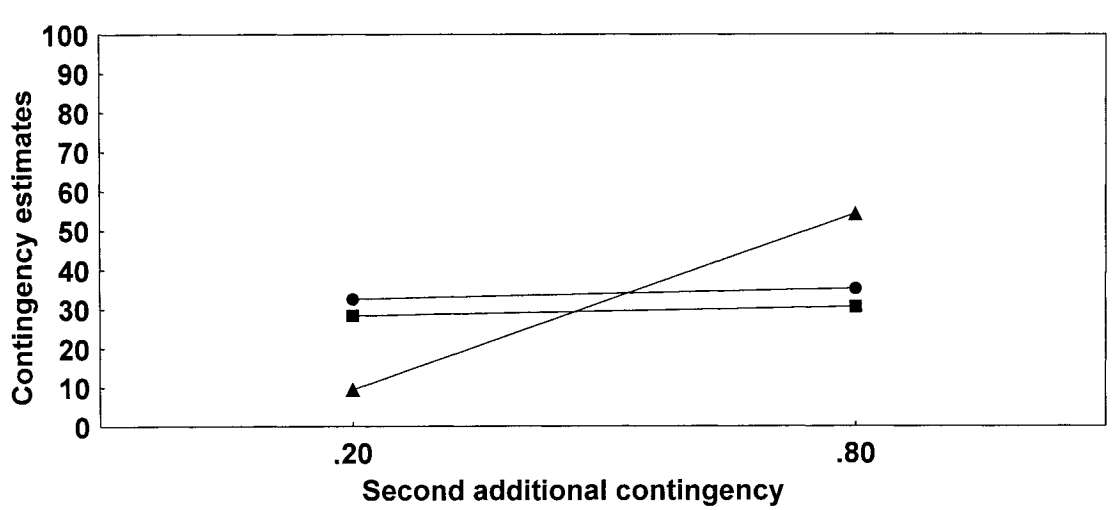

100
90
80
70
60
50
40
30
20
10
0
Confidence
4
6
8
40
Sample size
DP .80
DP .60
DP .40
DP .20

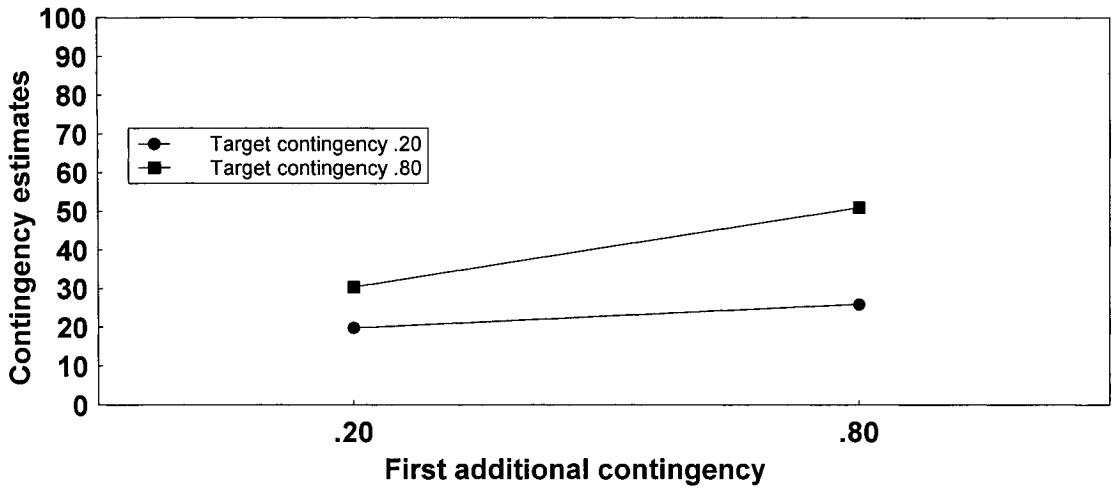

Médicament/Medication A
Maladie/Disease 1
Médicament/Medication B
Maladie/Disease 2
Médicament/Medication C
Maladie/Disease 3

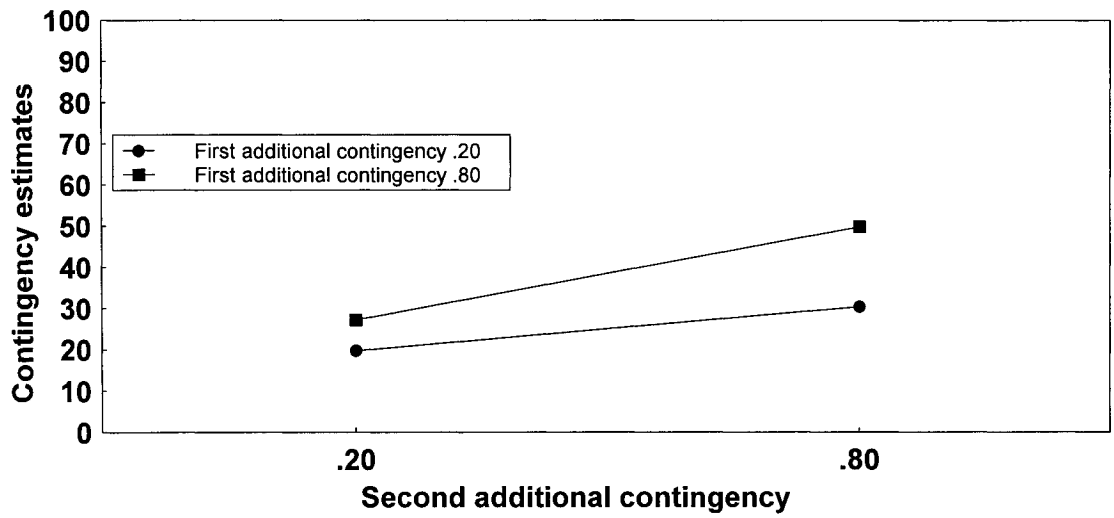

100
90
80
70
60
50
40
30
20
10
0
Contingency estimates
.20
.80
Target contingency
One contingency
Two contingencies
Three contingencies

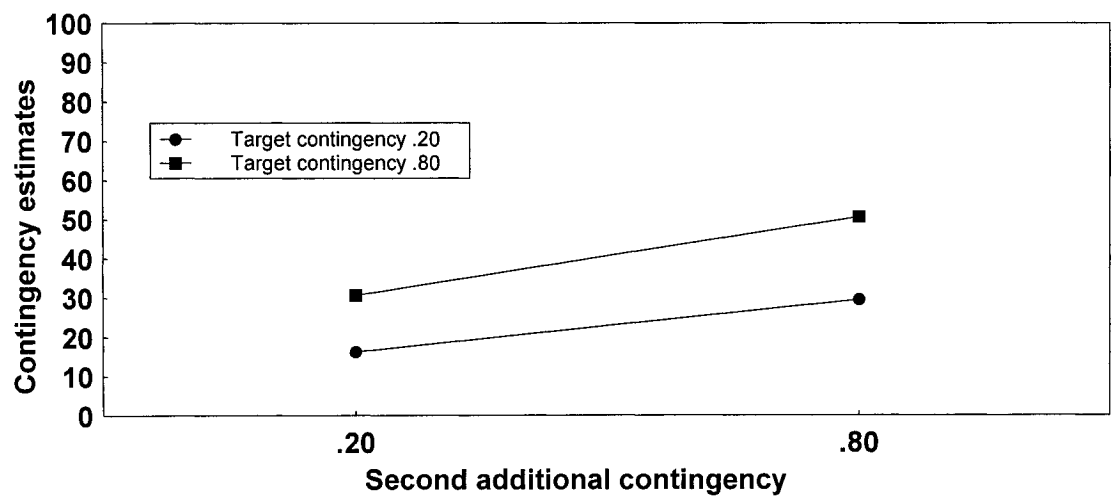

Target contingency
First additional contingency
Second additional contingency
Contingency estimates
100
90
80
70
60
50
40
30
20
10
0
.20
.80
Target contingency

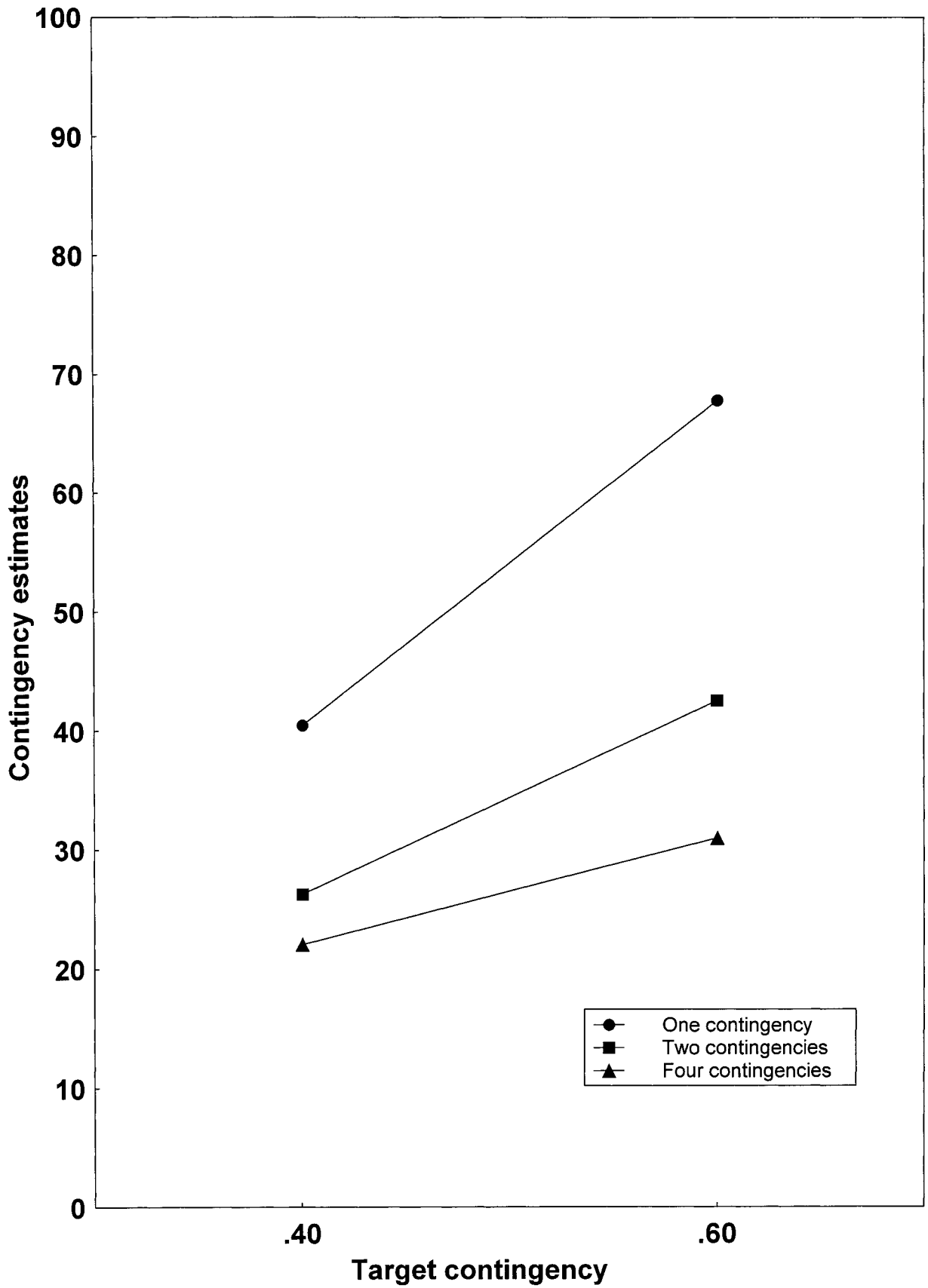

Contingency estimates
100
90
80
70
60
50
40
30
20
10
0
.20
.80
First additional contingency

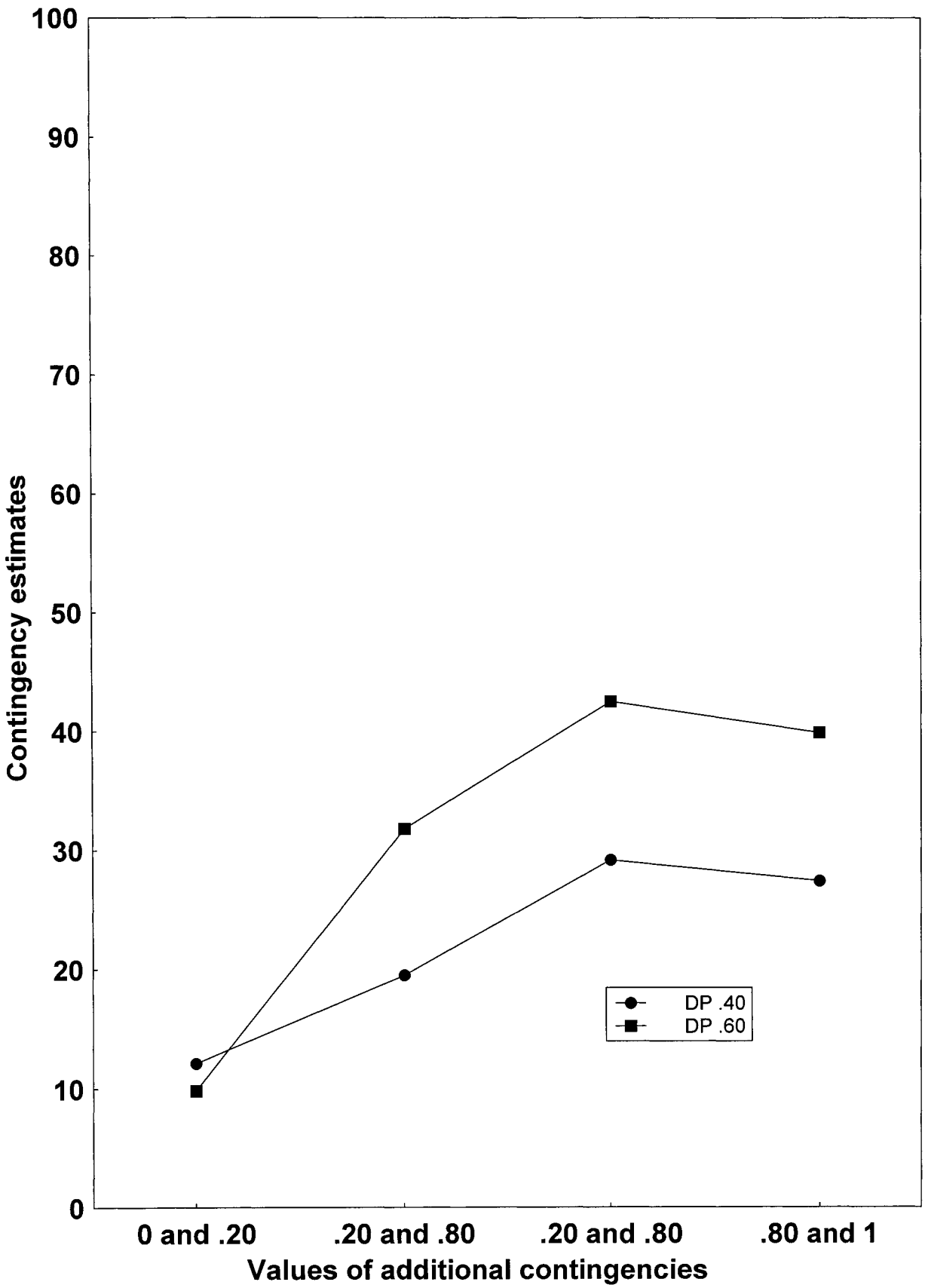

Contingency estimates
100
90
80
70
60
50
40
30
20
10
0
.20
.80
Second additional contingency

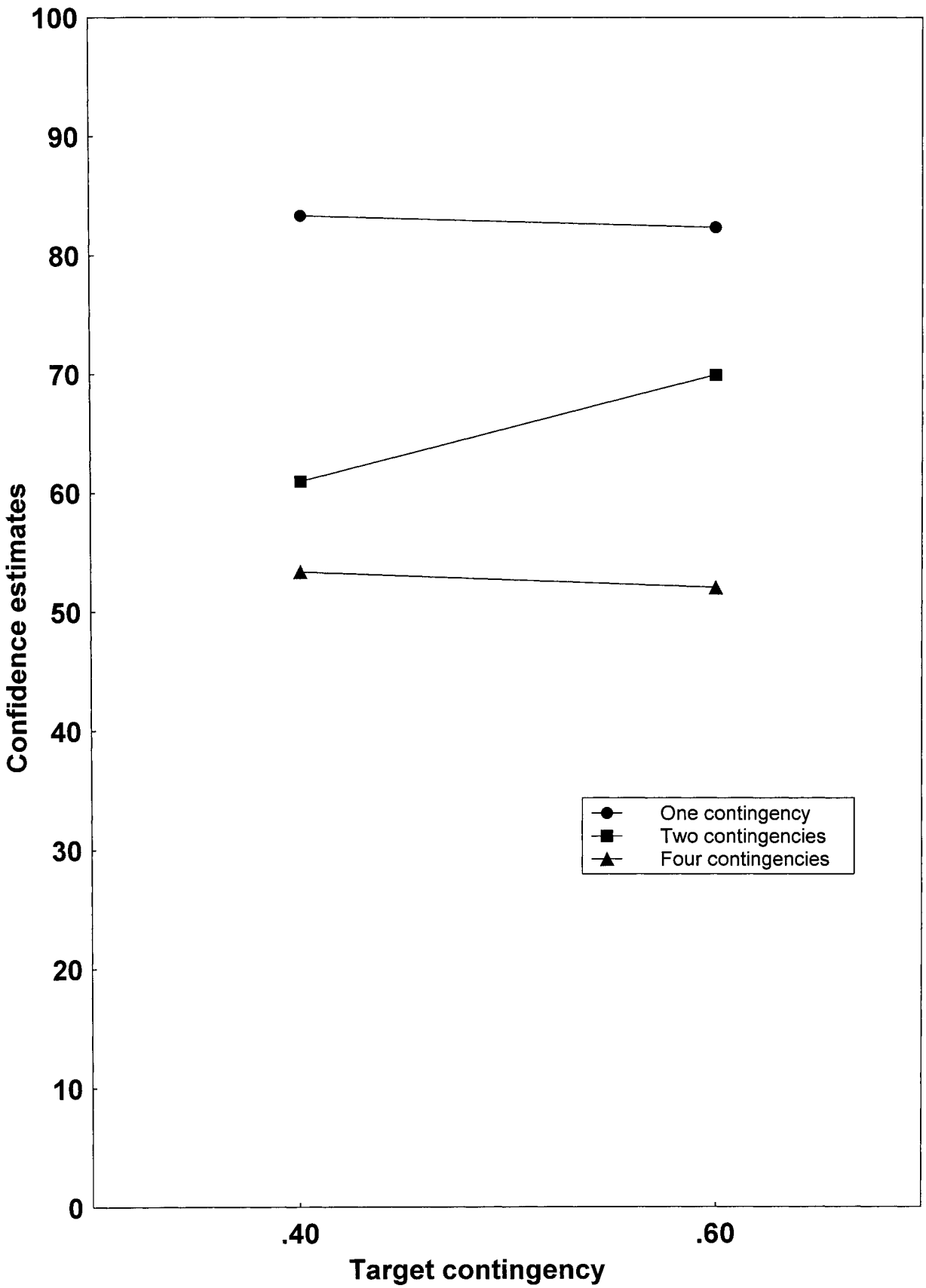

Contingency estimates
100
90
80
70
60
50
40
30
20
10
0
Target contingency .20
Target contingency .80
.20
.80
First additional contingency

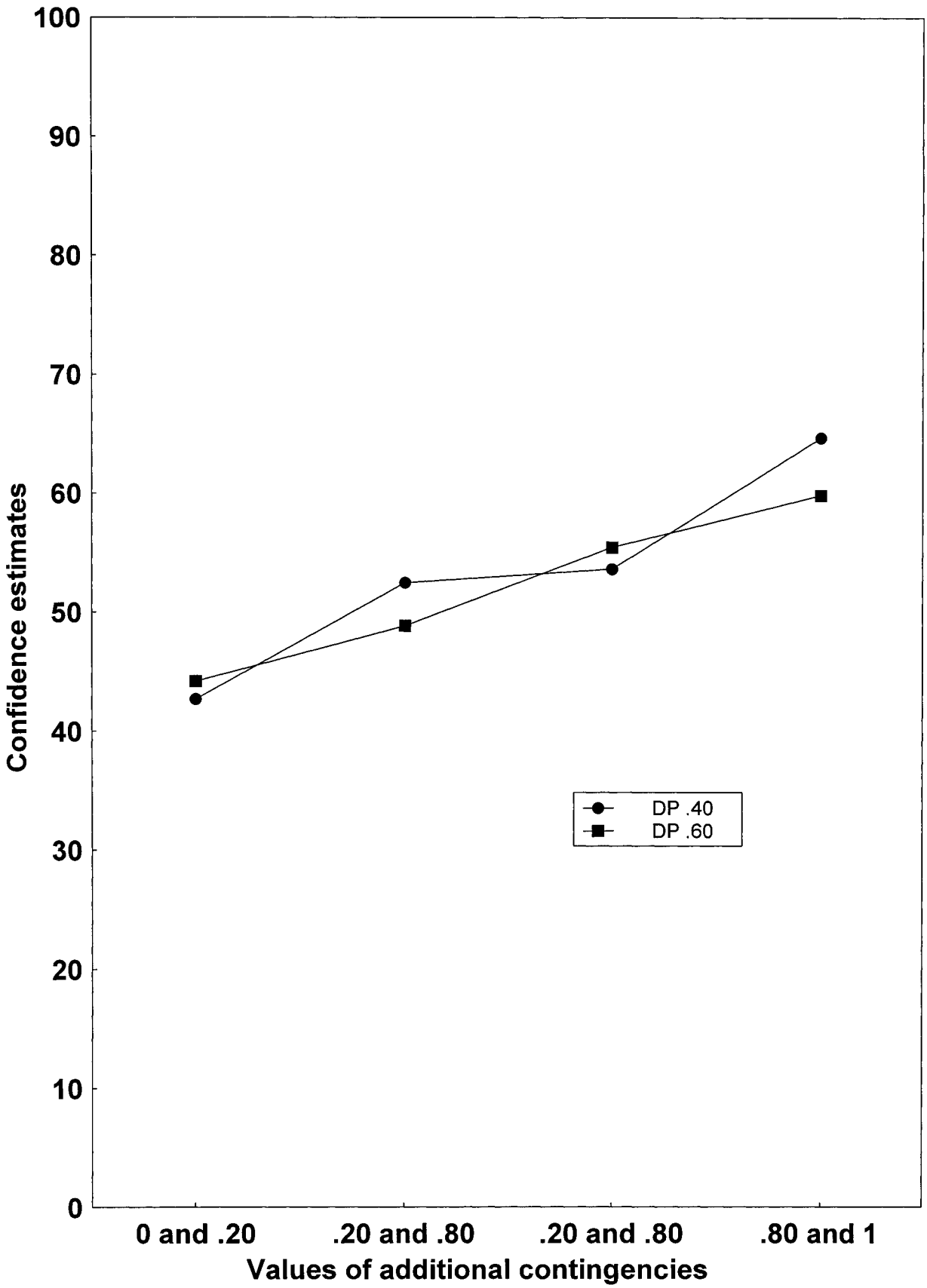

Contingency estimates
100
90
80
70
60
50
40
30
20
10
0
First additional contingency .20
First additional contingency .80
.20
.80
Second additional contingency

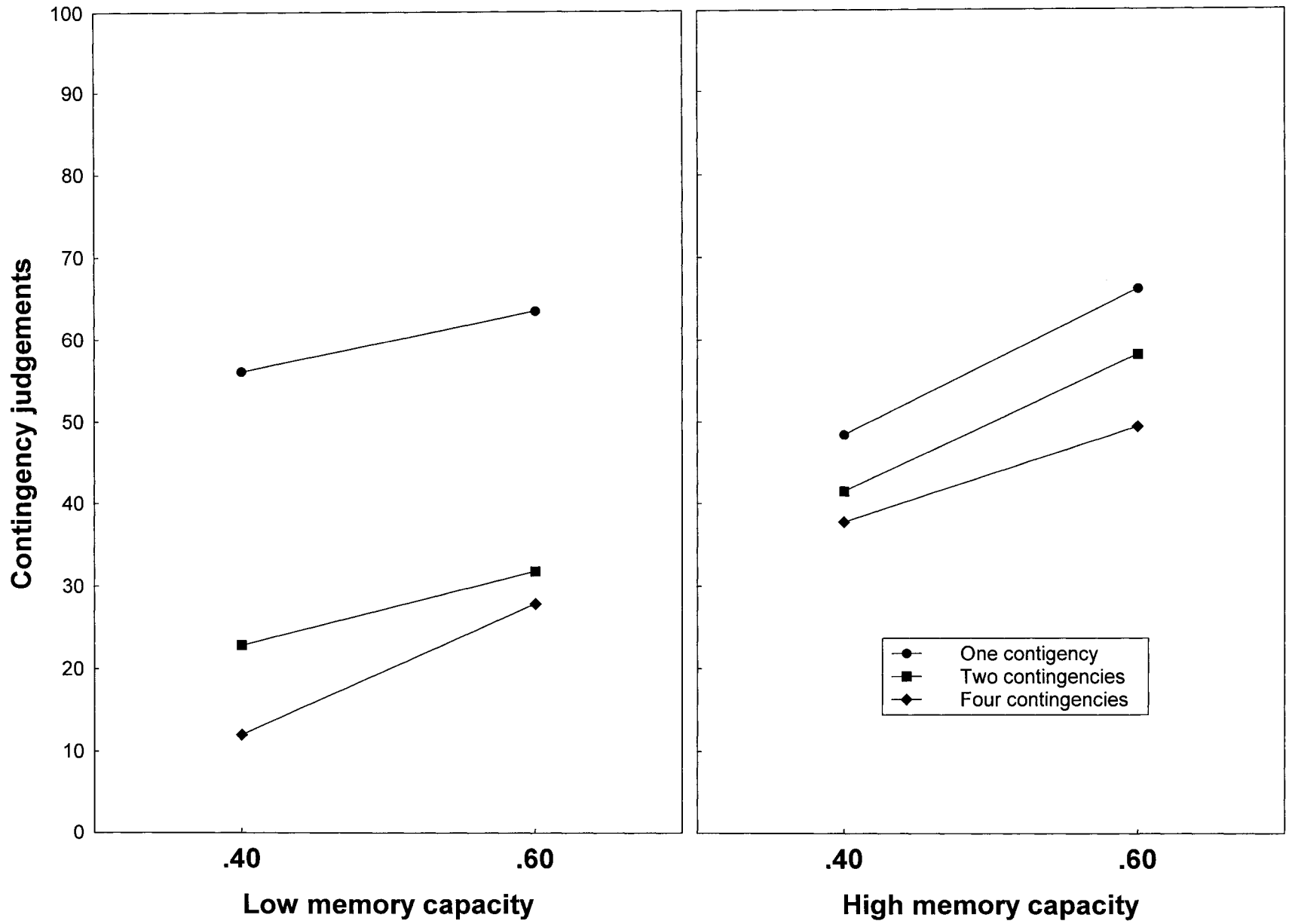

Contingency estimates
100
90
80
70
60
50
40
30
20
10
0
Target contingency .20
Target contingency .80
.20
.80
Second additional contingency

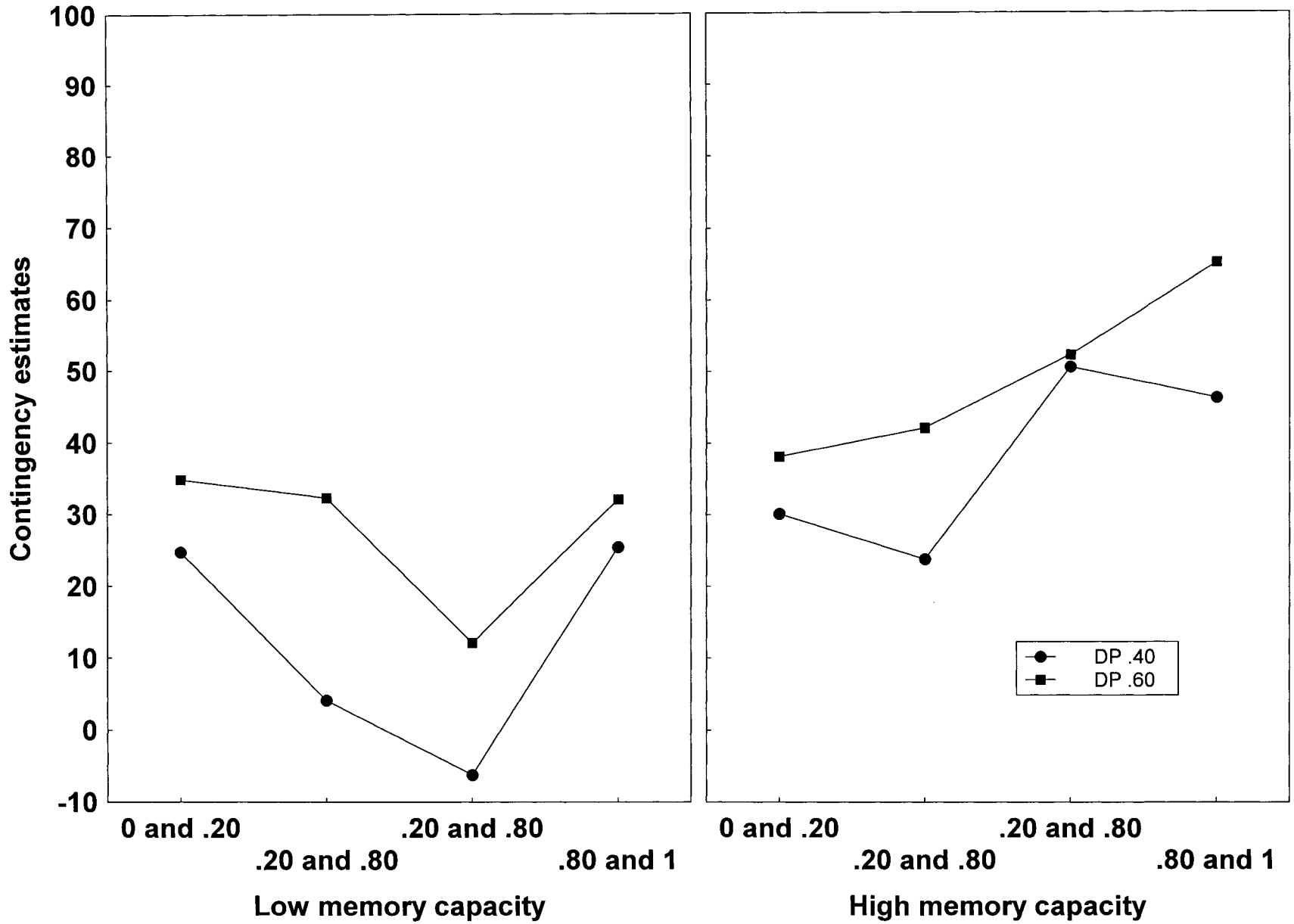

100
90
80
70
60
50
40
30
20
10
0
Contingency estimates
.40
.60
Target contingency
One contingency
Two contingencies
Four contingencies

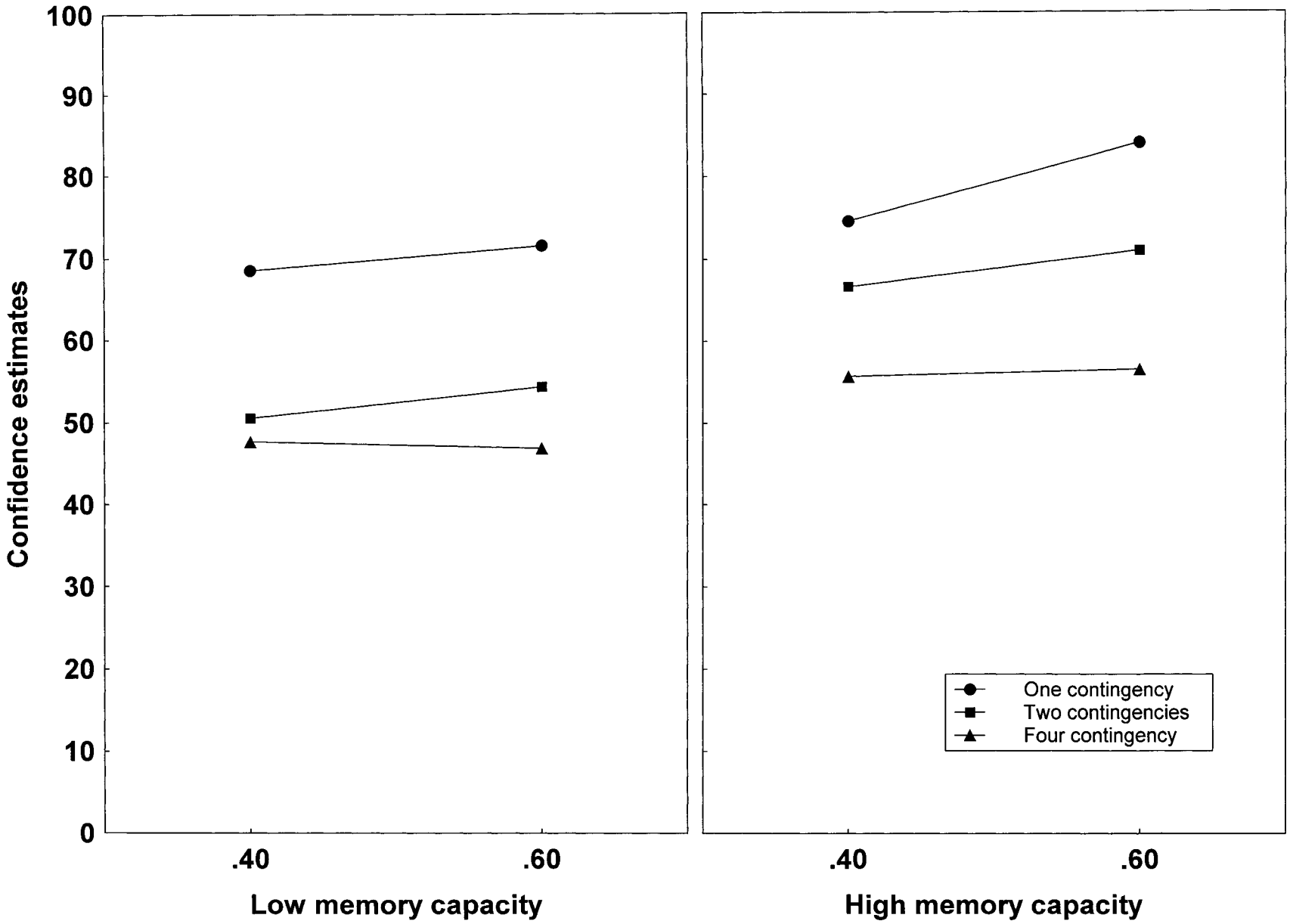

100
90
80
70
60
50
40
30
20
10
0
Contingency estimates
DP .40
DP .60
0 and .20
.20 and .80
.20 and .80
.80 and 1
Values of additional contingencies

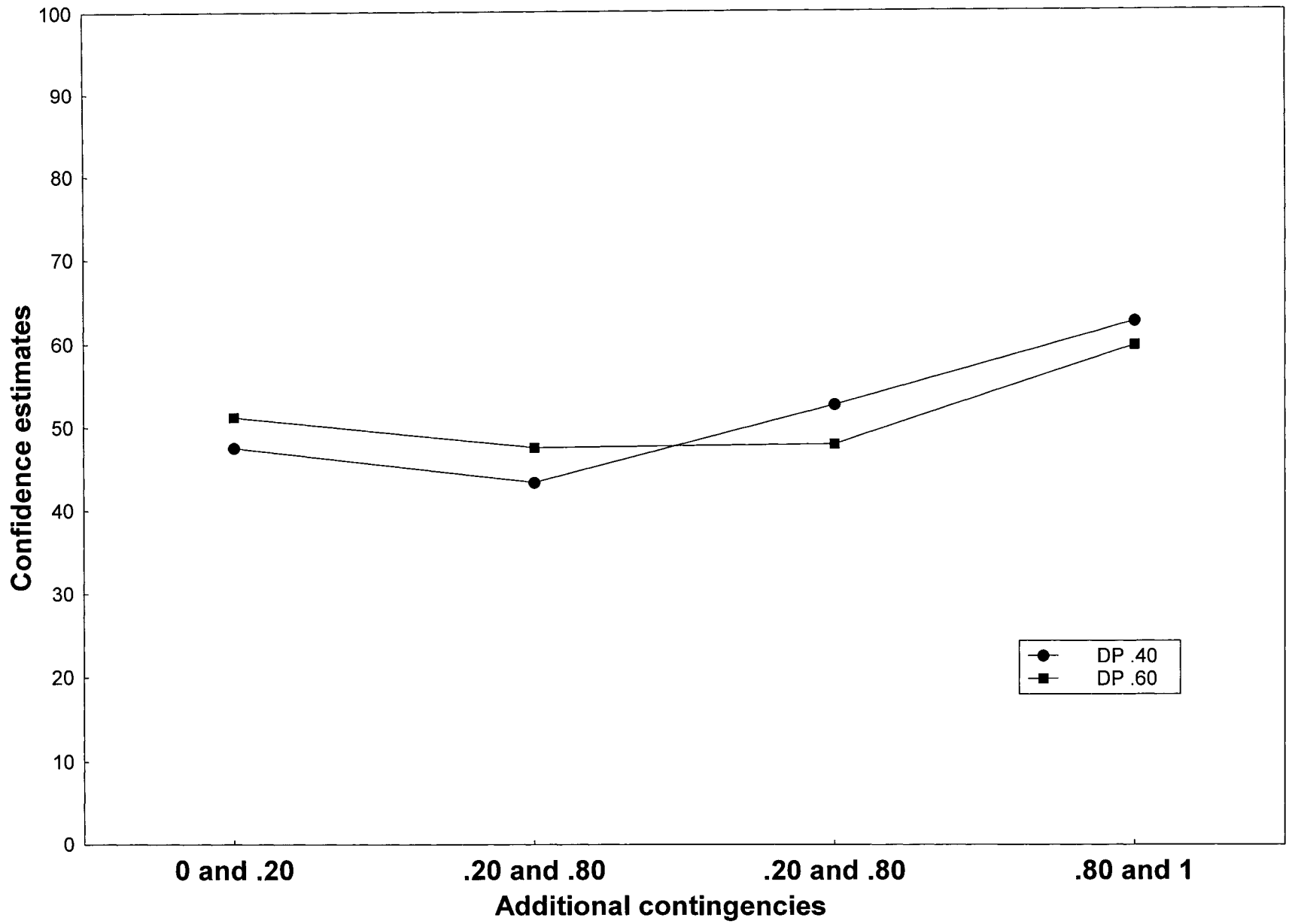

100
90
80
70
60
50
40
30
20
10
0
Confidence estimates
One contingency
Two contingencies
Four contingencies
.40
.60
Target contingency

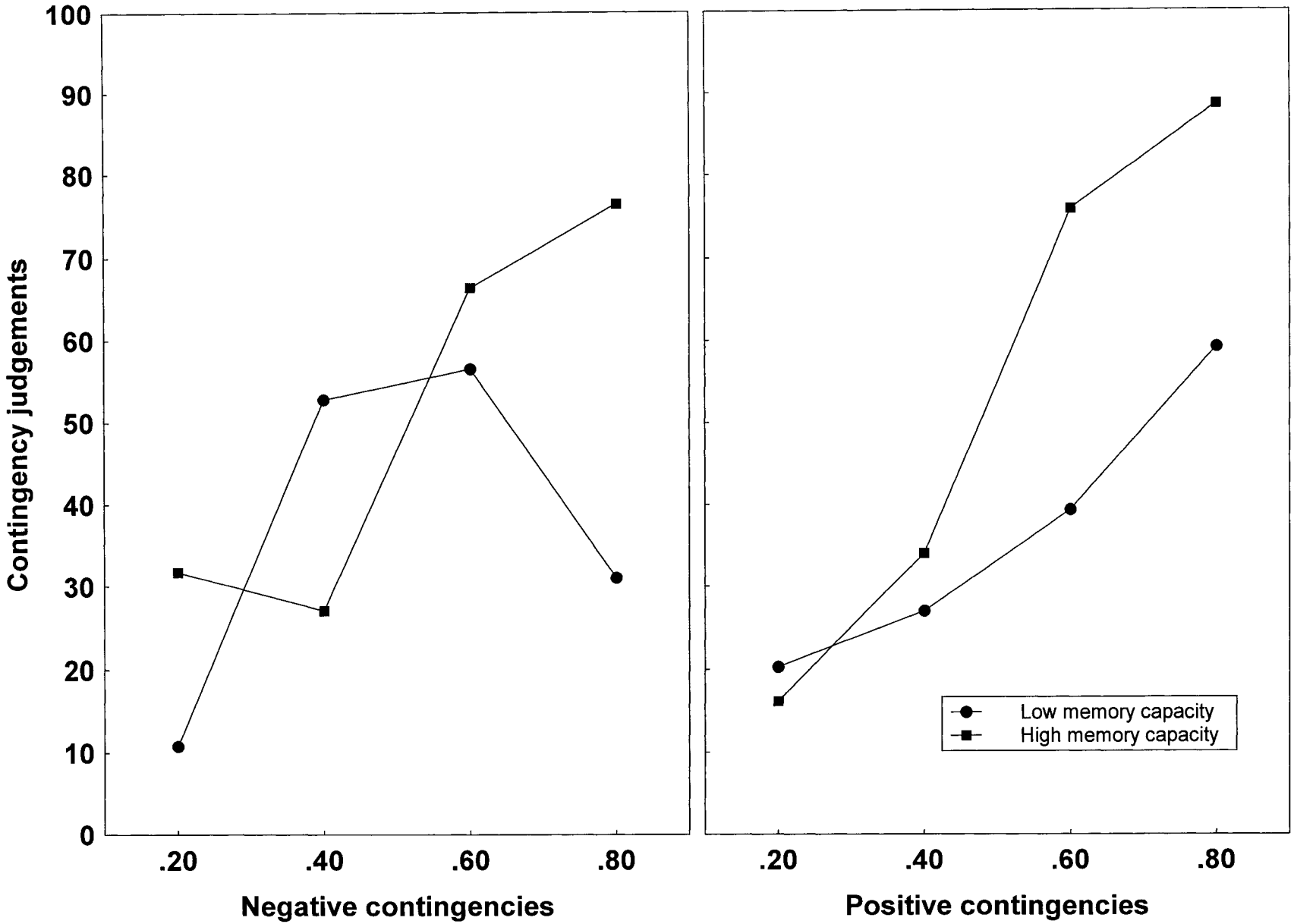

100
90
80
70
60
50
40
30
20
10
0
Confidence estimates
DP .40
DP .60
0 and .20
.20 and .80
.20 and .80
.80 and 1
Values of additional contingencies

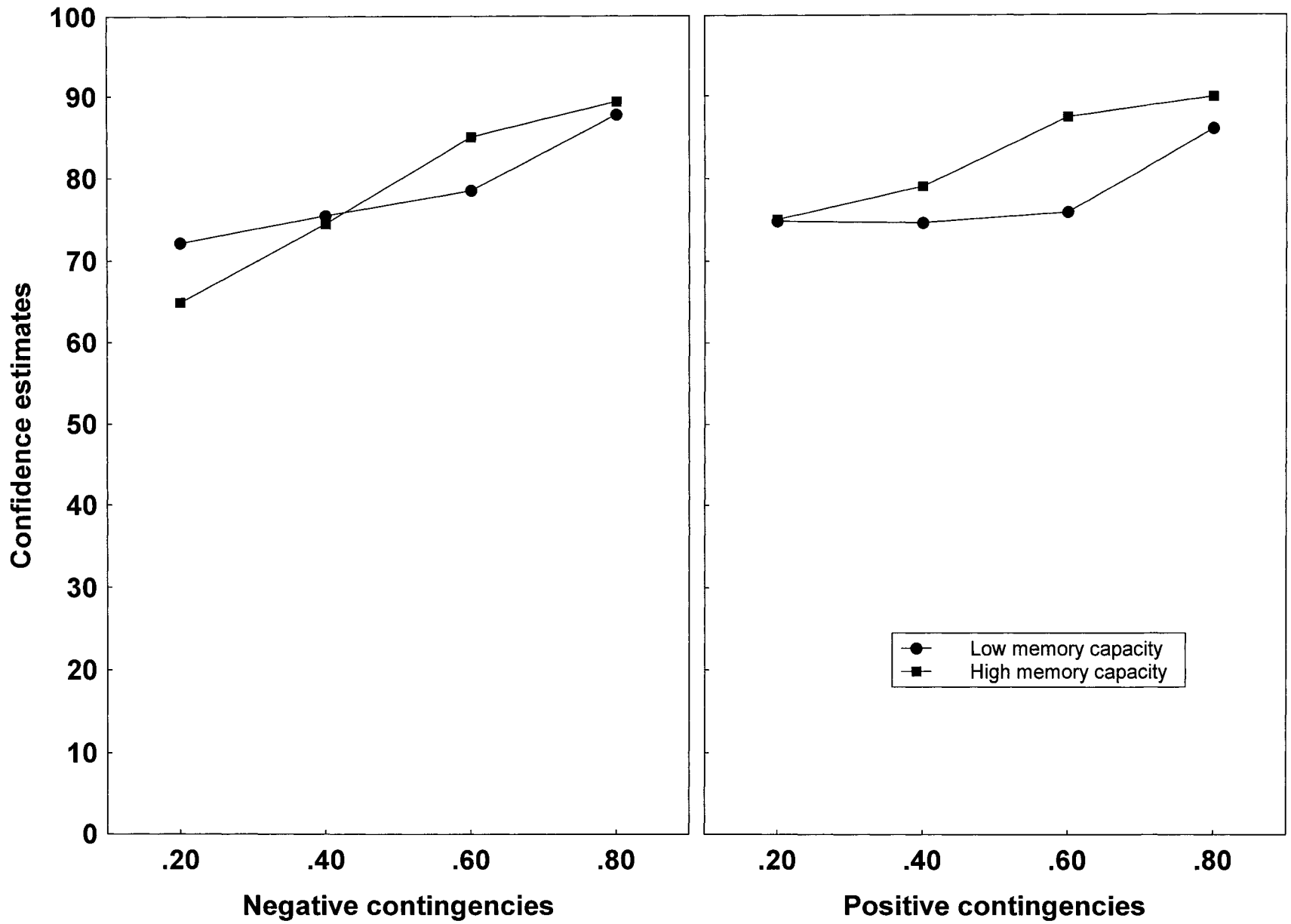

Contingency judgements
100
90
80
70
60
50
40
30
20
10
0
.40
.60
Low memory capacity
.40
.60
High memory capacity
One contigency
Two contingencies
Four contingencies


100
90
80
70
60
50
40
30
20
10
0
-10
Contingency estimates
0 and .20
.20 and .80
.20 and .80
.80 and 1
Low memory capacity
0 and .20
.20 and .80
.20 and .80
.80 and 1
High memory capacity
DP .40
DP .60


100
90
80
70
60
50
40
30
20
10
0
Confidence estimates
.40
.60
Low memory capacity
.40
.60
High memory capacity
One contingency
Two contingencies
Four contingency


100
90
80
70
60
50
40
30
20
10
0
Confidence estimates
DP .40
DP .60
0 and .20
.20 and .80
.20 and .80
.80 and 1
Additional contingencies


100
90
80
70
60
50
40
30
20
10
0
Contingency judgements
.20
.40
.60
.80
Negative contingencies
.20
.40
.60
.80
Positive contingencies
Low memory capacity
High memory capacity


100
90
80
70
60
50
40
30
20
10
0
Confidence estimates
.20
.40
.60
.80
Negative contingencies
.20
.40
.60
.80
Positive contingencies
Low memory capacity
High memory capacity