Winter 2005

The effect of memory for serially presented causal information on judgments of contingency

Christopher A. Barnes
University of New Hampshire, Durham

THE EFFECT OF MEMORY FOR SERIALLY PRESENTED CAUSAL INFORMATION ON JUDGMENTS OF CONTINGENCY

BY

CHRISTOPHER A. BARNES

Bachelors of Science, Morgan State University, 2000

Masters of Arts, University of New Hampshire, 2002

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirement for the Degree of

Doctor of Philosophy

in

Psychology

December, 2005

UMI Number: 3198004

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This dissertation has been examined and approved.

Dissertation Director, Victor A. Benassi, Professor of Psychology

Ellen S. Cohn, Professor of Psychology

Gary S. Goldstein, Associate Professor of Psychology

Michael J. Middleton, Assistant Professor of Education

Edward J. O'Brien, Professor of Psychology

August 17th 2005

Date

DEDICATION

This dissertation is dedicated to my family. I dedicate this to my wife, Angela, for serving as my strength in the past, present, and future. I could not have finished without your sacrifices. I also dedicate this to my first born son, Terrell, for serving as an endless source of motivation. I pray that our time apart during this process is the longest period that we spend not communicating with one another. Last, I dedicate this to my unborn; when you enter this world your father shall also be a man.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the priceless contributions of the following. I want to thank my advisor, Victor Benassi, for sharing his knowledge and patience. The greatest gift that you gave me was a desire to impress you with my writing. I admire you more than you know. I want to thank the entire dissertation committee for their support and comments that changed my way of thinking for the better: Ellen Cohn, Gary Goldstein, Michael Middleton, and Edward O'Brien. Edward O'Brien deserves a special mention for serving as my unofficial mentor and for organizing the subject pool---everyone in the Psychology Department should acknowledge you for the latter. I want to thank my research and teaching assistants that contributed to this project: Brendon Browne, Brendan Jorgensen, Gavin O'Brien, Lisa Tamagni, Chelsea Preston, and Jonathan Burg. Brendan Browne deserves special mention for his contributions from the start to the finish of this project. Last, I want to acknowledge the participants that arrived on time and the students in my classes that partook in the pilot studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

THE EFFECT OF MEMORY FOR SERIALLY PRESENTED CAUSAL INFORMATION ON JUDGMENTS OF CONTINGENCY

by

Christopher A. Barnes

University of New Hampshire, December, 2005

Four experiments investigated whether memory errors might account for errors in contingency judgments. Participants viewed contingencies one event at a time, later recalled the frequencies of the four event types, and judged the extent that they were related. Contingency judgments were more highly correlated with participants' memory of the contingency than with the actual contingency (Experiments 2 & 4); thus implying that inaccurate mental representations of the contingency contribute to erroneous judgments. Decreasing the time to view each event (i.e., from 3 to 5 s) increased the perceived difficulty of recalling event frequencies (Experiments 1 & 2), decreased the percentage of correct frequency estimations (Experiment 1), and increased the likelihood of a differential pattern of errors when recalling event frequencies (Experiment 1). Participants' knowledge of the total number of events (Experiments 2), their knowledge of the distribution of the four event types (Experiments 1-4), and the actual frequency of the event types were found to bias recalled event frequencies (i.e., in Experiments 3 and 4); the latter of which was also responsible for the differential pattern of errors when recalling event frequencies. In closing, the appropriateness of using a statistic calculated on one's memory of the contingency to assess judgment accuracy was discussed.

INTRODUCTION

Sensitivity to variable relatedness permits a variety of cognitive inferences. For example, an individual might foresee the occurrence of a future event after having observed a predictive event. Such foresight might permit the individual to delay, prevent, or prepare for the yet-to-occur event. Sensitivity to relatedness is especially useful in situations when the future event is potentially harmful. Upon viewing how events are related, an individual might make inferences about the cause of an event. Inferences due to sensitivity to variable relatedness have played a significant role in the lives of humans. Research has revealed that humans are sensitive to the direction and strength of relations (e.g., Allan & Jenkins, 1983; Ward & Jenkins, 1965; Wasserman & Shaklee, 1984). However, under many circumstances errors do occur when judging variable relatedness.

Many factors contribute to inaccurate judgments of variable relatedness (see Crocker, 1981, for a detailed discussion). Experimenters have primarily assessed whether individuals use accurate or error prone strategies to judge relatedness (e.g., Piaget & Inhelder, 1958; Shaklee, 1983; Ward & Jenkins, 1965). I argue here that memory errors contribute to inaccurate judgments of relatedness. Participants are able to distinguish positive, negative, and unrelated relations (e.g., Ward & Jenkins, 1965; Wasserman & Shaklee, 1984). The relevant information is the only way that a judge can successfully determine the strength and direction of a relation. Therefore, some representation of the relevant information is needed to form a veridical judgment. A representation is even more important when the individual must recall the relevant information from memory.

The experimental investigation follows a discussion of the research relevant to the role of memory for judging relatedness. First, the typical experimental paradigm will be discussed to highlight two procedures that differ in memory demands. Second, prior research will be reviewed that highlights crucial procedural factors that impact judgment competency. Third, three models will be discussed that can account for judgments of relatedness. The discussion of the models will be followed by an analysis of their assumptions pertaining to memory. Fourth, two areas of the research will be reviewed (i.e., subjective cell importance and order effects) which suggest that memory impacts judgments of relatedness. Last, the extant research relevant to memory and judgments of relatedness will be discussed.

The current review is limited to experiments that (1) require participants to judge the extent that two binary variables are related and (2) have implications relevant to the role of memory for judging relatedness. Several considerations influenced limiting the inclusion of experiments. First, the majority of prior research has investigated the extent that two binary variables are related. Second, the limitation minimizes the potential problem of comparing results obtained from different experimental paradigms (see Dennis & Ahn, 2002). Therefore, the limitation should facilitate an integrative discussion of experimental results. Last, no prior research has discussed the role of memory for judging relatedness using more than two binary variables.

The Experimental Paradigm

Several terms used throughout the literature must first be described. Event states refer to four forms of information, collectively referred to as causal information, that define the extent that variables are related. Causal information can be summarized in the cells of a 2 x 2 contingency table, as depicted in Figure B1 (labeled cells *a*, *b*, *c*, and *d*). The first of the two variables, Variable 1, is present in the cells on the top row (i.e., cells *a* and *b*) and absent in the cells on the bottom row (i.e., cells *c* and *d*). A similar rule holds for Variable 2 in the left (i.e., Variable 2 is present in cells *a* and *c*) and right columns (i.e., Variable 2 is absent in cells *b* and *d*). Cells *a* and *d*, collectively referred to as confirming cases, correspond to event states in which Variables 1 and 2 are either simultaneously present or absent and they serve to strengthen a positive relationship. Cells *b* and *c*, collectively referred to as disconfirming cases, correspond to event states in which only one of the variables is present at a time and they serve to strengthen a negative relationship. Each event state, and thus each cell of the contingency table, represents a unique combination of the presence and absence of the two variables.

The terminology that describes the variable relatedness must also be addressed. The extent that the two variables are related has been referred to as the relationship, relatedness, covariation, and contingency. Although each term is arguably interchangeable, the latter will be used throughout this discussion. Contingency also refers to the value of a statistic that quantifies the extent that two variables are related, referred to as *delta p* or Δp. *Delta p* is expressed by the following equation:

$$\Delta p = [\,A / (A + B)\,] - [\,C / (\,C + D\,)\,] \qquad (1)$$

where *A*, *B*, *C*, and *D* correspond to the frequencies of the similarly labeled event states (i.e., cells *a*, *b*, *c*, and *d*). The value of *delta p* is restricted to a range of +1.0 to -1.0. A contingency of +1.0 indicates that the presence of one variable perfectly predicts the presence of the other (i.e., a relation consisting only of *a* and *d* event states), while -1.0 indicates that the presence of one variable perfectly predicts the absence of the other (i.e., a relation that consisting only of *b* and *c* event states). A value of 0.0 indicates that no relationship exists between the two variables. Because *delta p* is calculated by to taking the difference between two conditional probabilities, a given contingency can be created using various combinations of conditional probabilities (or frequencies of causal information).

Researcher may choose from one of two procedures frequently used to present causal information. The two procedures, trial-by-trial (TBT) and summarized (SUM), differ in a variety of ways. The most important difference, at least for this discussion, pertains to the procedures' demands on memory.

The TBT procedure places a large burden on memory when judging contingency. Causal information is presented one event state at a time. Judgments are formed in the absence of all or the majority of the causal information. Variables 1 and 2 are often referred to as the input and output of the relation. Judges can passively observe the input and output (e.g., Jenkins & Ward, 1965) or actively initiate or withhold the input (i.e., the frequency of Variable 1) and observe its effect (i.e., the presence or absence of the output; e.g., Jenkins & Ward, 1965; Shanks, 1985; Wasserman & Shaklee, 1984). The active TBT procedure samples two probability generators to determine a trial's outcome if a response is initiated or withheld.

The SUM procedure greatly minimizes the burden on memory when judging contingency. The SUM procedure simultaneously displays all of the causal information. Judgments are formed in the presence of the summarized causal information, which greatly reduces and may even eliminate memory demands. Causal information has been summarized using a variety of methods, which includes a contingency table (e.g., Smedslund, 1963; Wasserman, Dorner, & Kao, 1990, Experiment 2), timeline (Wasserman & Shaklee, 1984), and statements that indicate the frequencies of each event state (e.g., Crocker, 1982; Wasserman, Dorner, & Kao, 1990, Experiments 1 & 3).

Researchers have investigated whether the TBT or the SUM procedure produces more accurate judgments. The consensus is that judgments are more accurate when using the SUM procedure (e.g., Ward & Jenkins, 1965; Wasserman & Shaklee, 1984). I argue that misremembering contributes to the poorer performance in the TBT procedure. The likelihood that the causal information is misremembered in the SUM procedure is greatly reduced because it is present when judging the contingency.

The dependent measures used to obtain judgments of relatedness have also differed. Instructions have requested, among other things, that participants judge the relationship (e.g., Maldonado, Catena, Candido, & Garcia, 1999), likelihood of future events (e.g., Shaklee & Tucker, 1983), control (e.g., Jenkins & Ward, 1965; Ward & Jenkins, 1965), influence and connection (e.g., Allan & Jenkins, 1980). Interest in terminology has faded because early research failed to demonstrate that it influenced contingency judgments (see Allan & Jenkins, 1983; Jenkins & Ward, 1965). The scales used to record judgments have either been unidirectional or bidirectional (see Allan, 1993). Unidirectional scales only provide information pertaining to the strength of the judgment. Therefore, positive and negative judgments can not be distinguished using unidirectional scales. Bidirectional scales provide information pertaining to both the strength and direction of the judgment. Numeric scales can be used to compare the value of the judgment to that of the contingency, which is not possible with nominal scales (e.g., less likely, more likely, etc.).

Early Research

Early research failed to demonstrate that humans produce veridical judgments of contingency (e.g., Inhelder & Piaget, 1958; Jenkins & Ward, 1965; Smedslund, 1963). Inhelder and Piaget (1958) conducted one of the first experiments that utilized two binary variables to investigate contingency judgment. Their primary focus was to assess the development of the concept of correlation in children. Interviews with the children, aged 12-15 years, suggested that they used a variety of erroneous rules to judge contingency. The older children were able to distinguish between confirming and disconfirming cases but their concepts of correlation were incomplete and frequently led to inaccurate judgments. The authors speculated that adults, due to a more complete concept of correlation, would produce more accurate judgments.

Inhelder and Piaget's (1958) speculation motivated the work of Smedslund (1963). Nursing students viewed cards that depicted either the presence or absence of a symptom and whether the patient was later diagnosed with a disease. Participants judged the contingency, estimated the frequencies of causal information, and rank ordered contingencies according to their strength. Participants performed poorly on all of the tasks. Experiment 2 removed peripheral cues from the cards and permitted participants more time to examine the causal information, which did not improve their performance.

Thus, another experiment suggested that humans are unable to form veridical judgments of contingency.

Jenkins and Ward (1965) also set out to determine whether individuals could accurately judge contingency. In three experiments, participants judged the extent that their responses (i.e., the active group) or those of another participant (i.e., the passive group) had on the occurrence of the output. The experiment was explained to participants as either a scoring task, in which responses could produce points, or a control task, in which they predicted a trial's outcome before responding. Judgments were inaccurate and uncorrelated with contingency in both the score and control conditions. The best predictor of participants' judgments, which also served as a bias, was the frequency of cell *a* event states (i.e., obtaining a "point" or producing the predicted outcome). Experiments 2 and 3 investigated whether participants had an improper understanding of control and improving their understanding of it would improve judgment competency. Explicit training reduced the bias associated with the frequency of cell *a* event states but failed to improve judgment accuracy.

Ward and Jenkins (1965) assessed whether different procedures impacted judgment competency. Contingencies were presented to groups using either the TBT, SUM, or both procedures. Participants that viewed the SUM procedure were more likely to produce veridical judgments than those that viewed the TBT procedure or both procedures. Their experiment demonstrated that judgment competency is improved if the SUM procedure is used (e.g., Jenkins & Ward, 1965; Smedslund, 1963), which suggested that conditions do exist in which humans can provide veridical judgments of contingency.

Researchers' failure to obtain veridical judgments of contingency was also influenced by variable symmetry. Variable symmetry refers to whether the binary states of either the input or output correspond to one or two variables. The two states of an asymmetric variable correspond to the presence or absence of a single variable, while those of a symmetric variable correspond to the presence of distinct events (i.e., blue or brown eyes). Because variable symmetry can be independently applied to the input (I) and output (O), it distinguishes four types of problem sets: 1I/1O, 1I/2O, 2I/1O, and 2I/2O.

Allan and Jenkins (1980; 1983) conducted an exhaustive investigation of the four types of problem sets defined by variable symmetry. Four aspects of their research prompted researchers to use asymmetric variables in later research. First, they concluded that variable symmetry contributed to the *not p-not q assumption* (Allan & Jenkins, 1980). The 2I/1O problem sets prohibited participants from observing the result of not responding (Allan & Jenkins, 1980), which led participants to erroneously conclude that no outcome will occur if no response is emitted. Therefore, participants used an inappropriate baseline when contrasting observations and forming their judgments of contingency. Second, they demonstrated that asymmetric variables are one way to mitigate the bias associated with the frequency of cell *a* event states (Allan & Jenkins, 1980; 1983). Third, Allan and Jenkins (1983) argued that if the input and output are not similarly defined by variable symmetry that the problem sets is not causally compatible. Biases were minimal when the input and output variables were similarly defined (i.e., 1I/1O or 2I/2O) and it was concluded that they were causally compatible. Causal incompatibility interfered with and biased participants' contingency judgments. Fourth, judgments were the most veridical with 1I/1O problem sets (Allan & Jenkins, 1983).

Explaining Contingency Judgment

The need for a theoretical account became apparent as the body of research investigating contingency judgments increased. Three models of contingency judgment will be discussed. Following their introductions, each theory's assumptions pertaining to memory will be discussed. The rule-analysis technique (Shaklee, 1983) will also be discussed immediately following the introduction of the first model.

Rule Based Model

Claims that judgments are formed according to a rule based strategies can be traced back to the first investigation of binary contingency judgments (Inhelder & Piaget, 1958). Although judgments may mimic a pattern predicted by a given rule, it does not eliminate the possibility that they are formed by another mechanism. This point is stressed because many researchers do not wish to imply that participants follow these rules when judging contingency (e.g., Wasserman, Elek, Chatlosh, & Baker, 1993).

Cell A Rule. The *cell a rule* states that participants form contingency judgments by comparing the number of *cell a* event states to those of the remaining cells. A relation is judged as positive if the frequency of *cell a* is the higher than each of the remaining cells, as negative if it is lower than each of the remaining cells, and as noncontingent if it is equal to each of the remaining cells.

A-Versus-B Rule. The *a-versus-b rule* compares the number of times the output variable occurs in the presence and absence of the input. In other words, it compares the frequency of cell *a* with that of cell *b* to distinguish among positive, negative, or noncontingent relations. A relation is judged as positive when *cell a* occurs more frequently than *cell b*, as negative when it occurs less than *cell b*, and as noncontingent when they are equal as frequent.

Use of these rules was thought to be an indication of immature reasoning. Neither the *cell a* or the *a-versus-b rule* requires a distinction between confirming and disconfirming cases. In addition, the *a-versus-b rule* does not rely on cell information from the entire contingency table. Use of a rule used all four cells and required distinguishing between confirming and disconfirming cases was thought to be an indication of mature reasoning (Inhelder & Piaget, 1958; Shaklee, 1983; Shaklee & Mims, 1982).

Sum-of-Diagonals Rule. The *sum-of-the-diagonals rule* incorporates all four cells and distinguishes between confirming and disconfirming cases (Inhelder & Piaget, 1958). The difference between the sums of the confirming and disconfirming cases, both its value and sign, are used to judge the relation. A relation is judged as positive if the difference is positive, as negative if the difference is negative, and as noncontingent if the difference is zero.

Conditional Probabilities Rule. The *conditional probabilities rule* or *delta rule* is the only rule that always leads to the correct strength and direction (Ward & Jenkins, 1965). The equation for the *delta rule* is the same as that used to define contingency (See Equation 1). The *delta rule* takes the difference between two conditional probabilities: (1) the probability of the output in the presence of the input (i.e., $A/ [A + B]$) and (2) the probability of the output in the absence of the input (i.e., $C/ [C + D]$). A relation is judged as positive if the difference is positive, as negative if the difference is negative, and as noncontingent if the difference is zero.

Rule-Analysis Technique. The rule-analysis technique (Shaklee, 1983) can determine which a participant used to judge contingency. The rule-analysis technique uses 12 contingencies that differ in the four rules success determining the direction of the contingency. The rule-analysis technique exploits the flaws the three lesser sophisticated rules to create four subsets of problems. The four subsets are created, one set for each rule, in a way that they can only be solved by a rule of equal or greater sophistication. The four rules, listed from the lowest to the highest level of sophistication, are as follows: *cell a*, *a-versus-b*, *sum-of-the-diagonals*, and *conditional probabilities*. The *cell a subset* can be correctly judged by all of the rules, the *a-versus-b subset* can only be correctly judged by the three more sophisticated rules, and so forth. Therefore, a judge's pattern of success across subsets can be used to determine the rule likely used to judge contingency.

Shaklee and colleagues have demonstrated the generality and reliability of the rule-analysis technique. The likelihood that children are categorized as having used a more sophisticated rules increases with age (Shaklee & Mims, 1981), which is consistent with that suggested by Inhelder and Piaget (1958). The rule-analysis technique also consistently assigns the majority of participants to moderately sophisticated rules (i.e., they use either the *sum-of-the-diagonals* or the *a-versus-b rule*; Shaklee & Hall, 1983; Shaklee & Mims, 1981; 1982; Shaklee & Tucker, 1980), which is consistent with individuals' frequent tendency to produce inaccurate judgments. However, it is possible that a participant's pattern of success is not consistent with the hierarchical arrangement of the problem sets. Such a participant might be using more than one rule or an unidentified rule. Inspections of participants' judgments that can not be categorized by the rule-analysis technique have not identified a new rule (Shaklee & Hall, 1983; Shaklee & Mims, 1981; 1982; Shaklee & Tucker, 1980).

The rule-analysis technique has several advantages over other techniques used to determine a participant's rule to judge contingency. First, it is not based on self-reports. Shaklee and Hall (1983) showed that self-reports, which have been used in the past (e.g., Smedslund, 1963), are not consistent with participants' success when judging the subsets. The one exception was that participants categorized as having used the *delta rule* provided explanations consistent with their patterns of success. Second, it is not based on the correlations between participants' judgments and that predicted from each of the rule-based strategies. Participants are said to have used the rule that predicts a pattern of success that is most highly correlated with the success of his or her judgments (Ward & Jenkins, 1965). This technique is problematic because the patterns of success of the four rules are highly correlated with one another and researchers often do not report the correlations between the other rules (Shaklee, 1983).

__Associative Model__

Three factors likely contributed to the use of the Associative Model to account for contingency judgments. First, researchers became increasingly reluctant to state that participants cognitively compute contingency judgments (e.g., Wasserman et al., 1993). Second, the Associative Model could easily account for cue competition effects which could not be explained by the existing models at that time (De Houwer & Beckers, 2002). Third, there are similarities between the results obtained in Pavlovian conditioning experiments and those in the contingency judgment literature (Alloy & Abramson, 1979;

Alloy & Tabachnik, 1984; Shanks, 1985; Shanks & Dickinson, 1987). The similarities become apparent if the input is considered equivalent to the conditioned stimulus (CS), the output as equivalent to the unconditioned stimulus (US), and the strength of the contingency judgment as equivalent to the strength of the conditioned response (De Houwer & Beckers, 2002; Shanks, 1985). Therefore, the Rescorla-Wagner model (Rescorla & Wagner, 1972), which explains Pavlovian conditioning, can also serve as a model of contingency judgment:

$$\Delta V_n = \alpha\beta\ (\lambda - \Sigma V_{n-1}) \qquad (2)$$

The model holds that ΔV_n is the change in the predictive strength that occurs on the current trial, α and β are learning-rate parameters that depend on the saliency of the input and output, respectively, λ is the maximum amount of predictive strength supported by the output, and ΣV_{n-1} is the algebraic sum of the predictive strengths of all stimuli that are present on each trial.

The model predicts a negatively accelerating learning curve for the increase in predictive strength (ΔV_n) of the input variable. Predictive strength will eventually reach an asymptote (λ), at which point judgments will be the most accurate and future changes in ΔV_n will be miniscule. The model suggests that stimuli (ΣV_{n-1}), which includes the context, actively compete for the limited amount of predictive strength (λ) on each trial (see the portion of Equation 2 in parentheses). Therefore, the predictive strength of the input variable only increases on trials when it is present. The competing stimuli obtain predictive strength when the input variable is absent. Because cue competition is a core feature of the Associative Model it can easily account for a number of retrospective revaluation effects. Retrospective revaluation effects refer to a variety of phenomena that alter the predictive strength of the input variable on trials in which it is absent. Retrospective revaluation effects will not be discussed here in detail because experiments typically include more than two variables and recent reviews are available elsewhere (Allan, 1993; De Houwer & Beckers, 2002; Shanks, 1993).

Belief Revision Model

The Belief Revision Model consists of two distinct mechanisms that collectively function as an anchor-and-adjust judgments of contingency (Catena, Maldonado, Candido, 1998; Catena, Maldonado, Megias, & Frese, 2002). The two mechanisms of the Belief Revision Model, the information-computing and information-integrating mechanisms, operate in a serial fashion.

The information-computing mechanism, which calculates the value of causal information, is activated first. Judgments of the causal information can be accounted for using the following equation:

$$\text{NewEvidence} = (w_1 * a + w_2 * b + w_3 * c + w_4 * d) / (a + b + c + d) \quad (3)$$

where a, b, c, and d again correspond to the frequencies of causal information and w_i corresponds to the subjective weights given to each form of causal information. Although the subjective weights are free parameters in the model, they are restricted to a range that is consistent with previous research (i.e., $|a| > |b| \geq |c| > |d|$; see Kao & Wasserman, 1993; Wasserman, Dorner, & Kao, 1990).

The information-integrating mechanism combines the unique contribution of the causal information with causal knowledge. Causal knowledge is assumed to be a function of the recently judged causal information as well as cognitive, motivational, and emotional factors. The information-integration mechanism incorporates NewEvidence using the following equation:

$$J_i = J_i\text{-}1 + \beta\,(\text{NewEvidence} - J_i\text{-}1) \quad (4)$$

where J_i corresponds to the resulting contingency judgment on the current trial, J_i-1 corresponds to relevant causal information and causal knowledge, and β corresponds to a learning rate parameter. The Belief Revision Model can account for order effects, which neither the Rule Based or Associative Models can account for, because of its anchoring information-integrating mechanism (e.g., Catena, Maldonado, & Candido, 1998; Catena, Maldonado, Megias, & Frese, 2002). However, the Belief Revision Model is limited in that it, like the Rule Based Model, cannot account for retrospective revaluation effects.

The Belief Revision Model is consistent with the intent of an integrative theory of contingency judgment proposed by Alloy and Tabachnik (1984). Their theory attempted to explain the interaction between prior expectations and causal information. The Belief Revision Model is superior to Alloy and Tabachnik's theory because it consists of a testable model instead of a theory driven review of the literature, which later received harsh criticism (Goddard & Allan, 1988).

<u>The Models and Memory</u>

The importance of memory for the models ranges from crucial to irrelevant (Baker, Murphy, & Vallee-Tourangeau, 1996). The following discussion highlights each theory's assumptions pertaining to memory.

The Rule Based Model assumes that contingency judgments rely heavily on the memory of causal information. Contingency judgments are a function of the rule used, the rule's inputs (i.e., causal information), and one's competency when computing the rule's output (i.e., a judgment). To arrive at a veridical judgment of contingency the judge must have perfect episodic memory of causal information (Baker, Murphy, & Vallee-Tourangeau, 1996). Therefore, failure to produce a judgment predicted by a rule-based strategy might reflect errors in estimating the frequencies of causal information.

The Associative Model assumes that episodic memory is of no importance to contingency judgments (Baker, Murphy, & Vallee-Tourangeau, 1996). Judgments are influenced by the empirical properties of the input and output (e.g., their temporal contiguity and proximity). Observing the empirical properties serve to strengthen their association and the resulting judgment is an indication of the strength of the association. Memory and an understanding of covariation are noticeably absent from the Associative Model because it was originally a model of animal conditioning (Alloy & Tabachnik, 1984). Researchers studying animal conditioning often challenged the existence of cognitive representations and would be extremely hesitant to assume that animals have them.

The Belief Revision Model assumes that memory serves two roles. First, the information-computing mechanism relies on memory of causal information in a manner similar to that described for the Rule Based Model. Second, the information-integrating mechanism relies on memory of causal knowledge (Maldonado et al., 1998). Judgments are the result of integrating causal knowledge and causal information. Memory of causal knowledge serves as the model's "anchor" and has the greatest impact on resulting judgments. Causal information also influences the resulting judgment but to a lesser extent; that is, assuming that the learning rate parameter is assigned a value greater than 0.0 but less than or equal to 1.0. The restriction on the learning parameter's range is necessary for any learning to occur (i.e., $\beta > 0.0$) and for the impact of the current causal information to "adjust", not dominate, the resulting judgment (i.e., $\beta \leq 1.0$).

Memory and Contingency Judgment

Few researchers have addressed the relationship between memory and contingency judgments. Two bodies of research are reviewed which suggests that memory impacts contingency judgments: subjective cell importance and serial order effects. In addition, several reviews of the literature are highlighted that have discussed the role of memory for judging contingency (e.g., Alloy & Abramson, 1979; Baker, Murphy, & Vallee-Tourangeau, 1996; Crocker, 1981; Wasserman, Elek, Chatlosh, & Baker, 1993). Last, experiments that have investigated the role of memory for judging contingency are reviewed (e.g., Shaklee & Mims, 1982; Yates & Curley, 1986).

Existing Bodies of Research

Two bodies of research suggests that memory plays a role in judging contingency. Research pertaining to subjective cell importance is discussed first and is followed by a discussion of order effects in contingency judgment. The former body of research is relevant to the encoding and recalling of causal information, while the later is relevant to the role of memory of past judgments.

Subjective Cell Importance. Research investigating subjective cell importance has many implications for the role of memory for judging contingency. Subjective cell importance can be defined as beliefs about the causal information necessary for judging contingency (Wasserman, Dorner, & Kao, 1990). *Delta p* is calculated on the frequencies of all four forms of causal information. However, participants do not similarly believe that all forms of causal information are equally important for judging contingency

(Crocker, 1982; Maldonado, Catena, Candid, & Garcia, 1999; Mandel & Lehman, 1998; Wasserman, Dorner, & Kao, 1990; Wasserman & Kao, 1993; White, 2003).

Crocker (1982) was the first to demonstrate that participants differentially weight the importance of causal information. Participants indicated which forms of information were both necessary and sufficient to judge contingency (e.g., the relation between practicing and the outcome of a tennis match). Participants had a strong tendency to rate the importance of causal information in the following order: $a > b > c > d$. However, the rank order switched when instructions explicitly requested what information was relevant for judging the likelihood of practicing and losing a match (i.e., $b > a > c > d$). Crocker discussed the importance of using unambiguous dependent measures. Wasserman, Dorner, and Kao (1990; Experiments 1 and 3) later replicated her finding and showed that the differential ratings holds when the order that participants rate the causal information is counterbalanced.

Wasserman, Dorner, and Kao (1990; Experiment 2) also sought to determine how causal information is used when judging contingency. A set of 25 contingency tables was constructed with combinations of three cell frequencies (i.e., 5, 10, or 20). The tables were constructed so that a set of three tables existed that only differed in the frequency of one cell. For example, a table defined by 10*a*, 10*b*, 10*c*, 10*d* could be compared to one defined by 15*a*, 10*b*, 10*c*, 10*d* and 20*a*, 10*b*, 10*c*, 10*d*. Therefore, any differences in judgment could be attributed to the use of cell *a*. The differential use of causal information was consistent with the differential ratings of cell importance (e.g., Crocker, 1982). Wasserman et al. (1990) concluded that participants must transform the values of causal information either prior to or while judging contingency.

Kao and Wasserman (1993) assessed the subjective weighting of cell information through two parameter fitting techniques. First, the weighting parameters for each cell could be allowed to vary while fitting them to participants' judgments. Therefore, one could assess which rule accounted for the most variance in participants' judgments. Second, restricting the range of parameters in accordance to the predictions of the four rules might help determine the rule used by participants. For example, the *a-versus-b rule* disregards cells *c* and *d* so their respective parameter weights can be assigned a value of 0.0. Participants would be said to have used the *a-versus-b rule* if their parameters for cells *a* and *b* each approximated .50. The two parameter fitting techniques were not successful predicting rule use. However, the resulting parameters did suggest a pattern of differential cell use consistent with the rank order of each cell's importance (e.g., Crocker, 1981; Wasserman, Dorner, & Kao, 1990).

Maldonado, Catena, Candid, and Garcia (1999, Experiments 4A & 4B) demonstrated that the subjective weights assigned to the event states are not fixed. Their rating scale differed from those used in previous experiments in that it was bidirectional (cf. Crocker, 1982; Wasserman, Dorner, & Kao, 1990), which permitted an indication of the magnitude and direction of each cells contribution to a judgment. The rank order of the absolute values replicated previous research (i.e., $|a| > |c| \geq |b| > |d|$; see Crocker, 1982; Wasserman, Dorner, & Kao, 1990); however, the ratings for cells *b* and *c* were negative in sign. Their experiment also indicated that preexposure to a noncontingent relation significantly reduced ratings of cell importance for cell *a*. Their finding is important because researchers did not previously entertain the concept of malleable cell weights.

The idea that subjective cell importance may influence memory of causal information has not been discussed in the literature. Researchers have primarily entertained the idea that participants do not equally weight the causal information when forming their judgments (Maldonado et al., 1999; Wasserman, Dorner, Kao, 1990). Crocker (1982) mentioned that participants might opt not to integrate recalled causal information due to beliefs about cell importance, but she did not entertain the notion that one's beliefs might influence memory. I argue that subjective cell importance might influence the encoding and/or recalling of causal information. Alloy, Crocker, and Tabachnik (1980; as cited by Alloy & Tabachnik, 1984) reported similar evidence when they demonstrated that prior expectations and biased information seeking can account for errors when recalling causal information. However, the question remains whether beliefs about the importance of cell information influences memory of it when no strong prior expectations exist.

Order Effects. Research investigating order effects is also relevant for the role of memory for judging contingency. Order effects can be defined as any reliable difference in contingency judgments that result solely from the order of presenting causal information. Order effects are important because the order of event states has no impact on the value of the resulting contingency. When using the TBT procedure, event states that can be arranged in multiple orders. Dennis and Ahn (2002) criticized experiments because they consisted of evenly distributed causal information, which fails to simulate real-life exposure. One interpretation of order effects is that individuals are more likely to forget causal information portrayed either at the beginning, middle, or end of a problem set. A second interpretation is that participants use memory of past judgments

(Maldonado, Catena, Candid, & Garcia, 1999) and causal information (e.g., Dennis & Ahn, 2002) as the reference point for forming contingency judgments.

Experiments investigating order effects in contingency judgment can be grouped into two categories. First, several experiments have demonstrated a primacy or recency effect (Dennis & Ahn, 2002; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996; Yates & Curley, 1986). A primacy effect is defined as the initial causal information having a greater impact on contingency judgments, while a recency effect is defined as the causal information experienced towards the end of a sequence having the greatest impact. Second, experiments have demonstrated that judgments are sensitive to the last event state encountered (Catena, Maldonado, & Candido, 1998; Catena, Maldonado, Megias, & Frese, 2002) and the type of contingency previously judged (i.e., a positive, negative, or noncontingent relation; Maldonado, Catena, Candid, & Garcia, 1999). Therefore, the current discussion will separately review experiments pertaining to memory effects (i.e., primacy and recency effects) and judgment sensitivity (cf. the interpretation of order effects proposed by Lopez, Shanks, Almaraz, & Fernandez, 1998).

Memory Effects. Yates and Curley (1986) found evidence of a primacy effect in contingency judgments. Participants judged a contingency composed of two blocks that differed only in their sign (i.e., Block 1 = .428; Block 2 = -.428). The two blocks were presented as one continuous sequence of 28 events using the TBT procedure. Half of the participants were told that they would later recall the cell frequencies to maintain a high level of attention throughout the problem set. Estimates of the two conditional probabilities required for the delta rule were used to determine contingency judgments, which showed evidence of a primacy effect in the forewarned and naïve conditions.

However, the primacy effect only reached statistical significance in the naive group. The authors concluded that forewarning participants of the recall task increased the likelihood that they attended to the entire problem set (see also Anderson & Hubert, 1963) and, as a result, reduced the magnitude of the primacy effect.

Dennis and Ahn (2001) investigated judgments of a noncontingent relation when the causal information was disproportionally distributed throughout the problem set. Participants judged a noncontingent relation that unevenly distributed an equal number of confirming and disconfirming cases. That is, 90% of the confirming cases were presented in either the first or second half of the problem set. Judgments were positive for participants that experienced the bulk of the confirming cases first and negative for those that experienced them in the second half. The authors claimed that a belief-updating model of contingency judgment could account for their data. They mentioned that the model should assume that participants disproportionally weight initial information, that initial information is used to form specific beliefs about the contingency, and that these beliefs are less influenced by causal information encountered after the specific beliefs are well formed. However, the Belief Revision Model was not mentioned as a suitable candidate or even cited in the experiment.

Judgment Sensitivity. Catena and colleagues (Catena, Maldonado, & Candido, 1998; Catena, Maldonado, Megias, & Frese, 2002) demonstrated that the frequency-of-judgment effect occurs when using a repetitive-judgment procedure[1]. The repetitive-judgment procedure requires that participants judge a problem set multiple times as more of the causal information is continually encountered. The frequency-of-judgment effect

[1] Catena and colleagues refer to the repetitive-judgment procedure as the trial-by-trial procedure. The name of the procedure was modified due to the term trial-by-trial referring to the method of presenting events, as is the norm in the literature.

refers to the tendency for participants' judgments to be sensitive to a confirming case if viewed prior to forming a judgment. The frequency-of-judgment effect has proven to be reliable, greatest after encountering a cell *a* event state, and has been demonstrated using summary tables (Catena et al., 2002, Experiment 2).

Catena and colleagues (i.e., Maldonado et al., 1999) also demonstrated that judging noncontingent relations affects later judgments of contingency. Judgments of positive and negative contingencies were lower when participants previously judged a noncontingent relation. The decrease in judgments for the negative contingency resulted in participants' judgments more closely approximating the actual contingency (thus replicating Maldonado, Martos, & Ramirez, 1991). However, a similar decrease in judgments for the positive contingency resulted in less accurate judgments. The researchers intended to replicate a similar phenomenon found in the animal literature referred to as learned irrelevance (e.g., Baker & MacKintosh, 1977; 1979), which is a retardation in the learning of relations upon preexposure to a noncontingent relation. However, the current experiment suggested that preexposure to a noncontingent relation facilitated the learning of a negative contingency. As a result, the authors argued that the Associative Model is flawed in its account of contingency judgment.

In conclusion, the research investigating order effects favors the second proposed interpretation. That is, that memory serves as the basis from which additional judgments are integrated. However, support for the second interpretation does not rule out the first. That is, it is still possible that the likelihood of forgetting causal information is also influenced by its position. The failure to support the first interpretation may be due be cause it is close to impossible to know exactly which event states are forgotten because repeat occurrences of causal information are indistinguishable. Therefore, it is not possible to test the first interpretation using the typical procedures used in contingency judgment experiments. Therefore, researchers might wish to assess memory of past judgments. Past judgments are thought to be integrated with causal knowledge and beliefs (Alloy & Tabachnik, 1984; Shanks, 1991; Maldonado et al., 1999), which might make it difficult to obtain a pure assessment of past judgments.

Reviews of the Role of Memory

Several reviews have addressed the role of memory for judging contingency (e.g., Baker et al., 1996; Crocker, 1982). The most comprehensive review is Crocker's (1982) six-step model of contingency judgment. Individuals are susceptible to biases that might contribute to inaccurate judgments at each step of Crocker's model: (Step 1) deciding what data are relevant, (Step 2) sampling cases, (Step 3) classifying instances, (Step 4) recalling the frequencies of causal information, (Step 5) integrating the evidence, and (Step 6) using the covariation estimate for future judgments. The following discussion will only review the steps that are most relevant to the role of memory for judging contingency.

First, Crocker (1982) argued that a judge must determine the appropriate forms of causal information (Step 1). She argued that participants actively seek out information that confirms the question asked, which might improve the likelihood that participants accurate recall the frequency of cell *a*. Seeking out cell *a* might improve memory of it due to the pursuit itself, intrinsic rewards, increased attention, or the increased likelihood of successful encoding. In summary, the process of determining the relevant information may affect the later recall of causal information.

Second, individuals must also correctly categorize causal information (Step 3). Correctly categorizing causal information might be particularly difficult for disconfirming cases, which are not as visually distinct as the confirming cases. Confirming cases, unlike the disconfirming cases, are easily distinguishable because the cues are either simultaneously present (cell *a*) or absent (cell *d*)[2]. As a result, disconfirming cases may be more likely to be erroneously encoded into memory. If disconfirming cases are erroneously categorized (i.e., 7a, 5b, 11c, 9d might be perceived instead as 7a, 7b, 9c, 9d), it would alter the perceived contingency (i.e., 0.03 versus 0.00, respectively)[3] and potentially influence the resulting judgment. Therefore, judgment errors might also be caused failures to correctly categorize causal information.

Third, individuals must recall all of the causal information when judging contingency (Step 4). Shaklee and Mims (1982) found evidence that the magnitude of errors when recalling cell frequencies increased as subjective cell importance decreased. (The magnitude of errors was as follows: $d > c > b > a$.) In addition, Crocker argued that causal information consistent with prior expectations might have an advantage of being successfully recalled (Crocker & Taylor, 1978; as cited by Alloy & Tabachnik, 1984). Therefore, judgment errors might be due to only a subset of the causal information being available when judging contingency.

---

[2] *Cell d* is represented by the absence of two variables when the output is asymmetrical.

[3] If the above described categorization error occurred, notice that it would differentially affect individuals that use different rule-based strategies to judge contingency. For example, *conditional probabilities rule* users should perceive a slightly more positive relation, *sum of diagonals rule* users should similarly perceive a noncontingent relation, *a-versus-b rule* users should instead perceive a negative relation, and it is unclear what type of judgment a *cell-a rule* user would perceive. The judgment of *cell-a rule* user can not be predicted because the remaining cells are both higher and lower in frequency. It could be argued that if a cell-a rule user were required to provide a judgment it would be done with very little confidence (See Alloy & Tabachnik, 1984, p.115, for a similar explanation for judgments when both situational and prior expectations are low).

Fourth, individuals will likely use the judgment for future use (Step 6). Memory of past judgments may deteriorate, change in direction, become exaggerated, or attenuated as time passes. The accuracy of past judgments is particularly relevant to the Belief Revision Model, which holds that it serves to anchor judgments of contingency (Maldonado et al., 1998). If past judgments are subject to forgetting or are altered as time passes, the resulting judgments would also be affected.

Assessments of Memory

Experimenters assessing the role of memory for judging contingency have done so either directly or indirectly. Direct assessments require participants to recall the frequencies of causal information, while indirect assessments require mental computations based on causal information. It is impossible to know whether errors in the indirect assessments are the result of faulty computations or discrepancies between the actual and used cell frequencies--that is, unless cell estimates are also recorded.

Direct Assessments. Shaklee and Mims (1982) conducted two experiments that required participants to recall cell frequencies. Their experiment was the first that reported detailed information about the recalled frequencies of causal information (cf., Smedslund, 1963; Yates & Curley, 1986). Previous experiments either did not distinguish between event states (Yates & Curley, 1986), only summarized the tendency to over- or underestimate cell frequencies (Smedslund, 1963), or did not require estimates of the entire contingency table (Arkes & Rothbart, 1985).

Shaklee and Mims (1982) required participants to recall the cell frequencies of 12 problem sets. In Experiment 1, one group recalled the cell frequencies and judged contingency. Estimates of the cell frequencies were not perfect. Judgments were recorded on a nominal scale which prevented an assessment of whether discrepancies between the actual and recalled contingency could account for errors in contingency judgments.

Indirect Assessments. Jenkins and Ward (1965, Experiment 2) indirectly addressed the role of memory for judging contingency. Participants estimated the conditional probabilities required to for the *delta rule*, which were used to calculate a subjective Δp (i.e., Δp_S). The estimated conditional probabilities were not reported and the resulting Δp_S values were not consistent with participants' judgments or with the contingencies.

Wasserman et al. (1993) also required that participants estimate the conditional probabilities, from which they calculated Δp_S. Wasserman et al. (1993) required that participants estimate the likelihood that their responses on a key produced or did not produce an outcome (i.e., light flash). The actual conditional probabilities derived from participants' responses were not consistent with their estimates. Judgments were reported to be more influenced by $[A/(A + B)]$ then by $[C/(C + D)]$. They argued that a weighting coefficient for $[C/(C + D)]$ could improve the fit but would be ad hoc and not warranted. I argue that the subjective cell importance literature validates such a modification. Their data were used as evidence against the Rule-Based Model and a lengthy discussion stated a desire to abandon its account of contingency judgment.

Summary

The current investigation aims to start a systematic investigation of the role of memory for judging contingency. Many explanations for erroneous judgments have been proposed in the literature. However, the possibility that memory contributes to inaccurate contingency judgments has received little attention. The argument for the role of memory for judging contingency consisted of a review of early research in the literature. Two existing bodies of research, subjective cell weighting and order effects, were also discussed. In addition, three models of contingency judgment and two variations of the typical procedure were discussed due to their relevance to memory demands when judging contingency. Last, the existing research pertaining to of the role of memory for judging contingency was discussed.

The following experimental investigation will consist of two phases. The first phase, Chapters I and II, aims to replicate and extend the results obtained by Shaklee and Mims (1982). The second phase, Chapters III and IV, aims to introduce a task distinction, popularized by Hastie and colleagues (Hastie & Park, 1986; Hastie & Pennington, 1989), to the contingency judgment literature. Its contents are dedicated to increasing our understanding of the role of memory for judging contingency.

CHAPTER I

EXPERIMENT 1: THE EFFECT OF SLIDE DURATION

Memory of the events that define a relation seems intuitively necessary to form a veridical judgment of contingency. Contingency (Δp) is the extent that two variables are related, which is mathematically expressed by:

$$\Delta p = [\, A \,/\, (A + B\,)\,] - [\, C \,/\, (\, C + D\,)\,] \qquad (1)$$

where *A*, *B*, *C*, and *D* are the frequencies of the event states. Event states, collectively referred to as causal information, are the four types of instances that can define the extent that two variables are related (See the cells labeled *a*, *b*, *c*, and *d* in Figure B1). Memory of causal information is especially important when judgments are formed in their absence, which is often the case. A discrepancy between contingency and a contingency judgment may be due, in part, to an inaccurate mental representation. That is, the judge may believe that the contingency is defined by a different set of events than it actually is. Consider an individual shown a problem set defined by: 7*a*, 5*b*, 7*c*, 5*d*. If he or she believes that a different set of event states was presented, their judgment would be expected to differ from that based on the actual set. The greater the discrepancy between the actual and perceived causal information, the greater the expected discrepancy between contingency and the judgment of contingency.

Many factors may result in a discrepancy between actual and perceived causal information. First, beliefs about the importance of causal information might bias encoding. Second, beliefs about the importance of causal information might bias its recall (Crocker, 1982). Third, participants might have difficulty distinguishing the four event states (Crocker, 1981). Last, the amount of causal information might exceed an individual's memory span. Each of the previously stated possibilities strongly suggests that the mental representation of causal information is not perfect.

The current investigation is not concerned with whether memory errors occur, but how they occur. A systematic pattern of errors might suggests one of two things: (1) that some forms of causal information are more salient than others and have a greater chance of being correctly recalled or (2) that beliefs about the importance of causal information influences their likelihood of being correctly recalled. It would also be particularly relevant if an inaccurate mental representation of causal information can account for judgment errors. If memory errors can account for judgment errors, judgments should be more consistent with the perceived contingency than with the actual contingency. The perceived contingency ($\Delta p'$), or inferred contingency, is the extent that the causal information stored in memory is related. The value of $\Delta p'$ is calculated using the same equation as Δp (see Equation 1); however, a participant's estimates of the frequencies of causal information are used instead of their actual frequencies.

Shaklee and Mims' (1982) procedure serves as a good starting point to begin the current investigation. In their experiment, participants recalled the frequencies of causal information that defined problem sets using a trial-by-trial (TBT) procedure. A TBT procedure presents causal information sequentially and one event state at a time. Both of their experiments demonstrated that the magnitude of errors increased across *cells a*, *b*, *c*, and *d*. However, the trend was not supported by significant differences in Experiment 2.

It is not possible to determine whether Shaklee and Mims' (1982) data are consistent with past research. The majority of past research has failed to report detailed analyses. For example, Smedslund (1963) and Crocker and Taylor (1978; as cited by Alloy & Tabachnik, 1984) reported only on the tendency to over- and underestimate cell frequencies. One experiment reported no consistent trend (Smedslund, 1963), while the other reported that judges overestimated information that was consistent with their expectations and underestimated information that was inconsistent with their expectations (Crocker & Taylor, 1978; as cited by Alloy & Tabachnik, 1984). Yates and Curley (1986) reported a statistical index (i.e., the sum of the deviations) that did not distinguish between event states. Alloy et al. (1980; as cited by Alloy & Tabachnik, 1984) never published their experiment indicating that biases in recall and information search can account for judgment errors, which prevents a detailed analysis it results. It is also likely that the sources of the errors in the Alloy's et al. (1980; as cited by Alloy & Tabachnik, 1984) and Shaklee and Mims' (1982) experiment are different. Alloy et al. (1980) assessed beliefs about the contingency itself, whereas the results reported by Shaklee and Mims' (1982) are consistent with beliefs about what information is important for judging contingency.

The limited research from which to compare Shaklee and Mims' (1982) results is problematic. A pilot study in my laboratory failed to replicate their results. However, the pilot study differed from their experiment in a number of ways. First, the pilot study used abstract cues to depict the input and output (i.e., Variables X and Y). Shaklee and Mims used contextual cues that depicted plausible relations, which may have activated specific beliefs (e.g., Alloy & Tabachnik, 1984; Shanks, 1991). Second, Shaklee and Mims (1982) required that participants view 12 problem sets. The pilot study only required that participants view one problem set. The pattern of errors may have emerged only after averaging data or judging several problem sets. Third, Shaklee and Mims (1982) used 24 event states to define problem sets. The pilot study only used 18 events states. A difference of 6 event states is small; however, its inclusion may have increased task difficulty and the need for a strategy to encode and/or recall causal information. Last, the pilot study did not require that participants judge the problem sets. Participants were only required to recall the cell frequencies. If the pattern of errors is due to beliefs about judging contingency, then the absent judgment task might explain why no pattern emerged in the pilot study. That is, participants may have viewed the pilot study as a memory task and equated the importance of causal information.

The proposed experiment more closely replicated the procedure used by Shaklee and Mims' (1982). Both abstract and contextual cues depict event states and the problem sets are identical to those used by Shaklee and Mims. Finally, cell estimates and judgments are required for each problem set. However, the current experiment also included a manipulation to impact the likelihood that strategies are used to aid memory. That is, the duration that each event state is shown varies across conditions. Event state duration was manipulated because statements from participants in the pilot study suggested that they adopted strategies to aid recall.

Several hypotheses are proposed in the current experiment. First, it is expected that the pattern of errors replicates that reported by Shaklee and Mims' (1982). That is, the magnitude of errors is expected to increase across *cells a*, *b*, *c*, and *d*. Second, it is expected that the magnitudes of errors are exaggerated in the condition with the shorter event state duration (i.e., 3 s). Third, participants exposed to the shorter event state duration are expected to perceive the longer event state duration as an easier task and vice versa. Last, it is expected that judgments are more highly correlated with $\Delta p'$ than with Δp.

Method

Participants

One hundred twenty students (27 males and 93 females, mean age = 18.71) enrolled in lower-level psychology courses at the University of New Hampshire participated in the current experiment. All participants received course credit for their participation. Of the participants, 70% were freshman and 10 % were upper classmen (i.e., 7.5% were juniors and 2.5% were seniors).

Design

A mixed design was used that consisted of a within-subjects factor (Cell) and three between-subjects factors (Block Order, Type of Cue, & Slide Duration).

Cell (4 levels), the within-subjects factor, refers to participants' estimates of the four forms of causal information. Estimates of causal information were subtracted from their actual frequencies to determine the absolute and signed cell deviations for each cell. Mean cell deviations, both absolute and signed, were calculated for each cell across the 12 problem sets. The Comparison Set was not included in the calculation of the means. The 12 problem sets were borrowed from Shaklee and Mims' (1982; Experiment 1); however, an oversight by the author resulted in one problem set failing to match that used by Shaklee and Mims (i.e., 8*a,* 5*b,* 9*c,* 2*d* was used instead of the intended 9*a,* 5*b,* 7*c,* 3*d*). Table A1 displays the actual frequencies of causal information and the Δp values for each problem set.

Twelve of the problem sets were arranged in three blocks of four. The order of the problem sets within blocks is listed in Table A1. Block Order (3 levels) refers to the order in which participants viewed the blocks: 1-2-3, 2-3-1, or 3-2-1. The Comparison Set was presented to all participants last, regardless of block order.

Type of Cue (2 levels) refers to whether the slides depicted abstract or contextual cues. In the abstract condition, cues defined the relation between Variables X and Y. In the context condition, cues defined the relation between plant watering and plant growth. The input and output in both conditions were asymmetric, meaning they corresponded to the presence or absence of a single event (Allan, 1993), and were displayed within a white box in the center of the slide. The input, if present, was depicted on the upper half of the rectangular box, while the output, if present, was depicted on its lower half. Cues in the abstract condition were typed in Times New Roman font, while cues in the context condition were hand drawings. Each slide contained peripheral cues that were unrelated to the contingency (The abstract condition contained a broken line that separated the top and bottom halves of the slide; the context condition contained the boot of the farmer that either watered or did not water the plant). Microsoft PowerPoint software displayed the slides via a PC compatible LCD projector.

Slide Duration (2 levels) refers to the event state duration. That is, the rate that slides advanced from one event state to the next. Event states either advanced at a rate of one slide every 3 s or one every 5 s. The 12 problem sets borrowed from Shaklee and

Mims (1982) used the same event state duration. However, the Comparison Set used the event state duration of the alternative condition.

Table A2 displays the number of participants in each experimental condition. Sessions were conducted in groups that ranged from 2 to 10 participants. Assignments to experimental conditions were random with the only restraint being that it did not result in a condition exceeding a total of 11 participants.

Measure and Items

Each measure contained a short description of the cues that defined the relation. The short description in the abstract condition read as follows:

> Some events occur as a result of another's presence. For example, observing Variable Y might be more likely when Variable X is also present. However, some events occur as a result of another's absence. For example, Variable Y might be more likely to occur when Variable X is absent. Still other events might be equally likely to occur whether or not another Variable is present. Use these slides to decide whether Variable Y is an event that is more likely to occur when Variable X is present, absent, or if it doesn't matter whether it is present or absent.

The short description in the context condition read as follows:

> Some plants need water to stay healthy. For example, their growth would depend on whether they received water each week. However, some plants need no water at all and will not grow if they are given water. These plants would be healthiest when given no water each week. Still other plants will grow regardless of whether they are watered or not. Use these slides to decide whether this plant is one that stays healthiest when watered, not watered, or if it doesn't matter whether they are watered each week.

Each measure required that participants estimate the frequencies of causal information, judge the relative likelihood of the output variable, judge the strength of the relation, and indicate their certainty in their previously completed judgments and cell estimates. There was also an item included only on the Comparison Set measure.

First, participants estimated the frequencies of causal information and entered their estimates into a blank contingency table. Below each table the following sentence could be read: *There are a total of 24 events in this problem set.*

Second, participants judged the relative likelihood of the output given the presence of the input. Judgments of relative likelihood were recorded on a scale ranging from +3 (*much more likely*) to -3 (*much less likely*). The following is an example of the relative likelihood item:

> For this relation, the slides indicate that the presence of the input resulted in the output being
>
> | +3 | +2 | +1 | 0 | -1 | -2 | -3 |
> |---|---|---|---|---|---|---|
> | much more likely | somewhat more likely | a bit more likely | just as likely | a bit less likely | somewhat less likely | much less likely |
>
> than when the input is absent. Circle the scale number that best completes the sentence.

The terms "relation", "input", and "output" were appropriately changed to match the cues used in the problem set.

Third, participants judged the strength of contingency. Judgments of strength were recorded on a unidirectional scale that ranged from 0 to ±100. Labels were located at the scale's 30, 70, midpoint (*50*), and two endpoints (*0* and ±*100*).

Fourth, participants reported their perceived certainty in the accuracy of their cell estimates and judgment of strength. Certainty ratings were recorded on a Likert Scale ranging from 1 (*just guessing*) to 10 (*absolutely certain*).

Last, participants indicated the relative difficulty when recalling the cell frequencies in the Comparison Set. That is, whether recalling cell frequencies was easier, harder, or just as easy/hard as the other problem sets. The scale ranged from +2 (*much easier than*) to -2 (*much harder than*) and read as follows:

After viewing the slides in this comparison set, recalling the number of instances were

+2	+1	0	-1	-2
much	somewhat	just as	somewhat	much
easier	easier	easy / hard	harder	harder
than	than	as	than	than

the previous sets completed in this experiment.

Procedure

Participants received a packet that included an informed consent form, abbreviated experimental instructions, debriefing form, and 13 multiple-item measures when they arrived. The experimenter began the session by reading the extended version of the experimental instructions, which read as follows:

> Our world is filled with things that go together, things that are influenced by one another, or things that relate to one another. We encounter these things in our everyday life. For instance, things may occur more frequently with one another. One example would be that taller people also tend to be heavier people. Things may also occur less frequently with one another. For instance, it is less likely to rain when it is currently sunny outside. Things may occur just as frequently in the presence or absence of something else. For instance, a car is just as likely to run out of gas whether it is painted red---or green---or whether you are driving alone or with a passenger. The paint job and passenger have no effect on running out of gas. In this experiment you will be given information about hypothetical events that may or may not go together---be influenced by one another---or relate to one another.
>
> For example, you may be asked the likelihood of it snowing and seeing a bird outside your window. Each time you look out your window, you may find that it is snowing and a bird is perched outside your window ---which will look like this {Show slide of cell *a*}. Notice that this information corresponds to Cell A, in the table in front of you, both the snow ---first row---and the bird---first column---are present. You might find that it is snowing, but there is no bird perched outside your window---which will look like this {Show slide of cell *b*}. Notice that this information corresponds to Cell B, the snow is present---it is still in the first row---, while the bird is absent---second column. You might also find that it is not snowing and that a bird is perched outside your window---which will look like this {Show slide of cell *c*}. Notice that this information corresponds to Cell C, the snow is absent---in the second row---, while the bird is present---first column. And lastly, you might find that it is not snowing and that there is no bird perched outside your window---which will look like this {Show slide of cell *d*}. Notice that this information corresponds to Cell D, both the snow---in the second row---and bird---second column---are absent.

After viewing 24 such instances, which may or may not be about snow and birds, you will be asked to recall the number of each type of instance. In the provided table similar to the one labeled with A, B, C and D, you will be asked to enter the number of times each of four event types occurred. In other words, you will be shown 24 events of any combination of the four possibilities just discussed. You will then be asked to recall from memory the number of times each of the four event types were observed---in other words---how many times you remember seeing each of the four events---and enter them into the blank table provided.

We ask that you put any writing utensils down while viewing the events. I think we all know that memory is not perfect, we wish to learn how memory works---so pay close attention and be sure to refer only to your memory---not that of your neighbors.

You will also be asked to {refer to the scale item} judge the relation between the two events. A value of positive three indicates that the events always occurred together---meaning that every time you looked out the window it was snowing and a bird was perched outside your window {Show slide of cell *a*}---or it wasn't snowing and no bird was perched outside your window {Show slide of cell *b*}---in other words, they were either always simultaneously present or absent---this would be a strict positive relationship. Snowing and seeing a bird were perfectly related.

A value of negative three indicates that the events never occurred together---meaning that every time you looked out the window it was snowing and no bird was perched outside your window {Show slide of cell *c*}---or it wasn't snowing and a bird was perched outside your window {Show slide of cell *d*}---in other words, they never simultaneously both occurred---this would be a strict negative relationship. Snowing and never seeing a bird were perfectly related.

A value of zero indicates that you were equally likely to see a bird outside your window whether it was or wasn't snowing---no relationship. Snowing had no effect on seeing a bird

As you might have guessed, we will not show you a strict relationships---that would be too easy---the relationships will be somewhere in the middle of the two extremes.

You will also be asked {refer to the scale item} to mark a numeric value on a scale ranging from 0 to 100 to describe the strength of the relation between the two events. Our primary interest with this item is your perceived strength of the relationship. Strict relationships come in two forms, both of which would receive a rating of 100. It may be that the events always occurred together. It may be that the events never occurred together. Both of which would receive a value of 100. If it didn't matter whether it was snowing, the rating should be zero. With this item we are not concerned with whether it is a positive or negative relationship---we simply want a measure of the strength of the relationship.

If you have any questions concerning the procedure described thus far, please ask them now. Remember---this is not a competition and you should not talk to, consult with or observe the work of others. And remember to place your pencils down when viewing each set of slides.

Hyphens indicated points at which the experimenter paused to facilitate participant's comprehension of the instructions.

The consent form was administered after the recital of the extended instructions and before presenting the first problem set. All participants choose to participate in the experiment.

The type of cues (abstract or context), order of blocks (Block Order 1-2-3, 2-3-1, or 3-1-2), and slide duration (3 or 5 s) all varied to define the 12 experimental conditions. Participants were given as much time as they needed to complete each measure and were told to look at the screen to inform the experimenter that they were ready for the next problem set. Upon completing the comparison Set, participants completed the set of demographic items. All participants were thanked for their participation, told to take the debriefing form, and were offered to stay after to receive an informal debriefing.

Results

Cell Deviations

The discrepancies between participants' estimates and the actual cell frequencies were used to create three distinct variables. First, absolute cell deviations were calculated to test the first hypotheses. Second, signed cell deviations were calculated to provide different, but equally important, information about participants' estimates. Mean absolute cell deviations provided information about the magnitude of errors, while mean signed cell deviations provided information about whether the magnitude of errors were greater for over- or underestimations. Errors due to over- and underestimations similarly affected the mean absolute cell deviations; however, they did not similarly affect the mean signed cell deviations because their respective signs were used to calculate them. Last, mean signed unit errors were calculated to provide information about the general tendency to over- or underestimate cell frequencies[4].

All data, means, and analyses were based on the 12 judged problem sets. Paired-samples *t* tests were used to determine whether differences existed between cells in the mean absolute and mean signed cell deviations. The planned set of *t* tests contrasted *cells a* and *b*, *cells b* and *c*, and *cells c* and *d.* In addition, tests were calculated on the data when collapsed across all conditions and separately for each slide duration condition.

Mean Absolute Cell Deviations. Mean absolute cell deviations are displayed in the top panel of Table A3. When collapsed all across conditions, the mean absolute cell deviation for *cell a* was significantly different than that obtained for *cell b*, $t(119) = -2.24$, $M_{SE} = 0.06$, $p < .05$. The remaining tests revealed no significant difference between *cells b* and *c* or between *cells c* and *d* (all $ps \geq .66$). Therefore, two additional follow-up *t* tests were used that contrasted *cells c* and *d* with *cell a.* The mean absolute cell deviation for *cell a* was significantly different than those obtained for *cells c* and *d*, *cell c*, $t(119) = -2.29$, $M_{SE} = 0.05$; *cell d*, $t(119) = -2.19$, $M_{SE} = 0.05$, all $ps < .05$.

The *t* tests indicated different patterns in the slide duration conditions. In the 3-s slide duration condition, the results were similar to that when the data were collapsed. The mean absolute cell deviation for *cell a* was significantly different than that obtained for *cell b*, $t(119) = -2.19$, $M_{SE} = 0.09$, $p < .05$ level. The remaining tests again revealed no significant difference between *cells b* and *c* or between *cells c* and *d* (all $ps \geq .28$).

---

[4] Participants' estimations of the actual cell frequencies were coded using the following conventions to calculate signed unit errors: overestimations were given a value of 1.0, underestimations were given a value of -1.0, and correct estimations were given a value of 0. The above described convention permitted an assessment of participants' tendency to overestimate or underestimate cell frequencies.

Therefore, two additional tests contrasted *cells c* and *d* with *cell a*. The mean absolute cell deviation for *cell a* was significantly different than those obtained for *cells c* and *d*, *cell c*, $t(119) = -2.79$, $M_{SE} = 0.09$; *cell d*, $t(119) = -1.91$, $M_{SE} = 0.08$, all $ps < .05$. In the 5-s slide duration condition, no significant differences existed between cells in any of the planned or follow-up *t* tests (all $ps \geq .10$).

The mean absolute cell deviations for the cells in the top row of the contingency table, *cells a* and *b*, were contrasted with those in the bottom row, *cells c* and *d*. Paired-samples *t* tests were conducted when collapsed across all conditions and separately for both slide durations; neither of which reached statistical significance (all $ps \geq .093$).

Mean Signed Cell Deviations. The mean signed cell deviations are displayed in the bottom panel of Table A3. When collapsed across all conditions, the mean signed cell deviation for *cell a* was significantly different than that obtained for *cell b*, $t(119) = -4.32$, $M_{SE} = 0.08$, $p < .05$. The remaining tests revealed no significant difference between *cells b* and *c* or between *cells c* and *d* (all $ps \geq .36$). Therefore, two additional follow-up tests were used that contrasted *cells c* and *d* with *cell a*. The mean signed cell deviation for *cell a* was significantly different than those obtained for *cells c* and *d*, *cell c*, $t(119) = 3.51$, $M_{SE} = 0.08$, $p < .05$; *cell d*, $t(119) = 2.55$, $M_{SE} = 0.09$, $p < .05$.

In the 3-s slide duration condition, the analysis revealed a similar pattern. The mean signed cell deviation for *cell a* was significantly greater than that obtained for *cell b*, $t(119) = 2.76$, $M_{SE} = 0.14$, $p < .05$ level. The remaining tests revealed no significant difference between *cells b* and *c* or between *cells c* and *d* (all $ps \geq .31$). Therefore, two follow-up tests contrasted *cells c* and *d* with *cell a*. The mean signed cell deviation for *cell a* was significantly different than that obtained for *cell c*, $t(119) = 2.93$, $M_{SE} = 0.96$, *p*

$< .05$. However, the mean signed cell deviation for *cell a* was not significantly different than that obtained for *cell d*, $t(119) = 1.71$, $M_{SE} = 1.10$, $p = .09$.

In the 5-s slide duration condition, the mean signed cell deviations were quite different. The mean signed cell deviation for *cell a* was significantly greater than that obtained for *cell b*, $t(119) = 3.53$, $M_{SE} = 0.08$, $p < .05$. Again, the remaining tests revealed no significant difference between *cells b* and *c* or between *cells c* and *d* (all $ps \geq .10$). However, the follow-up tests were not both significant. The mean signed cell deviation for *cell a* was not significantly different than that obtained for *cell c*, $t(119) = 1.95$, $M_{SE} = 0.96$, $p = .06$. Like the other analyses, the mean signed cell deviation for *cell a* was significantly different than that obtained for *cell d*, $t(119) = 2.02$ $M_{SE} = 0.10$, $p < .05$.

The mean signed cell deviation for the cells in the top row of the summary table were also contrasted with that for the cells of the bottom row. Paired samples *t* tests were conducted when the data were collapsed across all conditions and separately for both slide durations; neither of which reached statistical significance (all $ps \geq .13$).

<u>Signed Unit Errors</u>. Signed unit errors were calculated to determine the percentage of estimates that were over-, under-, and correct estimations (Table A4). The mean percentages of correct estimations were calculated across the four forms of causal information. An independent samples *t* test indicated that the difference in the mean percentages of correct estimations in the two slide duration conditions was significant, $t(6) = -9.64$, $Diff_{SE} = 1.85$, $p < .01$.

Mean signed unit errors are also displayed in Table A4. Mean signed unit errors can range from -1.0 to +1.0. Positive values indicate that errors were more likely to be overestimations, while negative values indicate that errors were more likely to be underestimations. There were no hypotheses or planned comparisons for the mean signed unit errors. Therefore, the data will be reviewed in the Discussion.

Multivariate Analyses

A multivariate analyses of variance (MANOVA) determined whether there was any effect of Slide Duration (2 levels), Type of Cue (2 levels), and/or Block Order (3 levels) on the mean absolute cell deviations. Slide Duration had a significant main effect on all four forms of causal information, *cell a,* $F(1) = 6.84$, $MS_e = 3.83$, $p < .05$; *cell b,* $F(1) = 14.37$, $MS_e = 7.02$, $p < .05$; *cell c,* $F(1) = 15.91$, $MS_e = 11.07$, $p < .05$; and *cell d,* $F(1) = 8.94$, $MS_e = 6.34$, $p < .05$. However, neither Type of Cue nor Block Order had a significant main effect on any of the mean absolute cell deviations. In addition, no interaction with any combination of factors reached statistical significance.

Relative Difficulty in Recall

An independent-samples *t* test was conducted on the relative difficulty item to determine whether differences existed in the two slide duration conditions. The difference between the 3-s and 5-s slide duration conditions approached but did not reach conventional significance, $t(6) = 1.79$, $Diff_{SE} = .51$, $p = .08$.

Judgments

Rule-Analysis Technique. Problem sets were designed so that one's success judging their directions can determine the rule used to judge contingency. That is, a participant's pattern of success judging contingencies can be used to facilitate the *rule-analysis technique.* A participant was said to have correctly judged a problem set if they correctly determined its direction. Meaning that a participant chose one of the positive options when judging a positive contingency (i.e., "*much more likely*", "*somewhat more* *likely*", or "*a bit more likely*"), one of the negative options when judging a negative contingency (i.e., "*much less likely*", "*somewhat less likely*", or "*a bit less likely*"), or "just as likely" when judging a noncontingent relation.

The rule-analysis technique categorized participants as having used a rule-based strategy if the following requirements were fulfilled: (1) at least two of the three problems from a rule's subset were correctly judged (i.e., *cell a subset*; *a versus b subset*; *sum of diagonals subset*; and *conditional probabilities subset*) and (2) at least two of the three problems of all subsets of lesser sophistication were also correctly judged. Participants were categorized as having used the rule of the greatest sophistication of which the previously mentioned requirements were met.

The design of the rule-analysis technique (see Shaklee, 1983 for a detailed description) makes it unlikely that a participant using one of the lesser sophisticated rules can successfully judge a contingency from a more sophisticated subset[5]. Participants whose success judging subsets does not conform to the design of the rule-analysis technique may have done one of two things: (1) they may have used multiple rules to judge contingency or (2) they may have used an unidentified rule. It is impossible to determine which possibility applies to such a participant; therefore, they are categorized as *unclassifiable*. Participants who do not correctly solve at least two problems from any of the subsets are categorized as having used *strategy 0*.

The percentages of participants categorized as having used the various rules are displayed in Table A5. There were no hypotheses or planned comparisons proposed for the rule-analysis technique. Therefore, the data will be reviewed in the Discussion.

---

[5] It is even less likely that a participant using a rule of lower sophistication can correctly judge two of the three problems from a rule's subset that is higher in sophistication.

Judgments. Pearsons correlation coefficients (r) were calculated to determine the relation between judgments of strength, Δp, and $\Delta p'$ (Table A6). The judgment of strength scale was unidirectional, which made it more appropriate to use the absolute values of Δp and $\Delta p'$ to calculate r values. When data were collapsed across all conditions, judgments of strength were significantly correlated with Δp, $r = .06$, $n = 1420$, $p < .05$. Judgments of strength were also significantly correlated with $\Delta p'$, $r = .06$, $n = 1415$, $p < .05$. The difference between the two previously mentioned correlations was minuscule. The strongest relationship, by far, was the relationship between Δp and $\Delta p'$, $r = .83$, $n = 1431$, $p < .05$.

Discussion

Cell Deviations

Mean Absolute Cell Deviations. To test the first hypothesis, a set of planned paired-samples t tests were performed on the mean absolute cell deviations. The general trend across cells was not consistent with that reported by Shaklee and Mims' (1982; i.e., the magnitude of errors did not increase across *cells a*, *b*, *c*, and *d*) in any of the reported analyses (see Table A3). Therefore, the planned tests did not adequately characterize differences in the mean absolute cell deviations and, as a result, additional follow-up tests were conducted.

The follow-up tests indicated that the magnitude of the errors for *cell a* was significantly less than those for the remaining three cells. This trend occurred in all but the analyses conducted on the 5-s slide duration condition, which suggested that it was a reliable effect. The lack of any significant differences in the 5-s slide duration condition suggests that participants may have adopted strategies to aid recall. Participants also might have adopted strategies in the 3-s slide duration condition; however, time might not have permitted efficient use. Unfortunately, the strategy that participants may have used, if at all, can only be speculated.

Mean Signed Cell Deviations. The mean signed cell deviations revealed that the magnitude of errors for *cell a* was greatest for overestimations. The exact opposite was true for the remaining cells. This trend resulted in the mean signed cell deviations for *cell a* being, for the most part, significantly greater than those for the remaining three cells. It may be that *cell a*, the co-occurrence of two variables, is a more salient cue that results in its frequency being consistently overestimated. A tendency to overestimate *cell a* might also account for the tendency to underestimate the remaining cells. The experimental instruction clearly stated that problem sets were defined by 24 events. Therefore, an effort to ensure that estimates totaled 24 events would force an overestimation of *cell a* result in at least one underestimation of the remaining cells.

Signed Unit Errors. The signed unit errors helped determine whether differences existed in the likelihood of correctly estimating cell frequencies in the two slide duration conditions. The independent-samples *t* test indicated that the percentage of correct estimations were significantly greater in the 5-s slide duration ($M = 68.23$, $SD = 1.30$) condition than in the 3-s slide duration condition ($M = 50.43$, $SD = 3.46$). The signed unit errors also indicated that, for the most part, more than 50% of the cell estimates for each cell were correct. Two exceptions occurred in the condition with the shorter event state duration for *cells c* (46.9%) and *d* (48.4 %).

The mean signed unit errors were also examined because its sign can be used as an index of the tendency to over- or underestimate cell frequencies. Participants were more likely to overestimate *cell a* and underestimate the remaining cells in all but the 5-s slide duration condition (Table A4). The signs of the mean signed unit errors and mean signed cell deviations were consistent with one another; the one exception occurred in the 5-s slide duration condition for *cell d* (*M*s = .03 and -.02, respectively; *SD*s = .69 and 56, respectively). The reported trend in the mean signed unit errors differed from Smedslund's (1963) experiment; in which, no trend existed. However, the data reported by Smedslund (1963) were not averaged across multiple problem sets like they were in the current experiment.

Multivariate Analyses

A MANOVA was conducted on the mean absolute cell deviations to test the second hypothesis. Of particular interest was the main effect of Slide Duration. Type of Cue and Block Order were also assessed; however, no differences were expected and their inclusion served as manipulation checks. The MANOVA indicated that only Slide Duration had a significant main effect on any of the four mean absolute cell deviations. As expected, mean absolute cell deviations were higher in the 3-s slide duration condition (Table A3). Block Order and Type of Cue served as manipulation checks; therefore, their lack of significance is an indication that they did not have an effect on the mean absolute cell deviations.

Relative Difficulty of Recall

To test the third hypothesis, responses on the relative difficulty item were compared in the two slide duration conditions. The difference in participant's perceived difficulty recalling cell frequencies was in the predicted direction; however, the difference failed to reach statistical significance. Participants indicated that it was easier to recall cell frequencies when given more time to view event states ($M = .25$, $SD = .80$) and that it was harder to recall the cell frequencies given less time ($M = -.02$, $SD = .79$). This topic will be discussed further in the Conclusions.

Judgments

Rule-Analysis Technique. The percentage of participants categorized as having used each rule is in many ways similar to that reported by Shaklee and Mims (1982). First, Shaklee and Mims (1982; Experiment 2) found, as did we, that about a third of the participants (i.e., 27 % when collapsed across all conditions of their experiment) were *unclassifiable*. This trend proved to be true in the current experiment when the data were collapsed across all conditions and in each slide duration condition (3-sec: 37.7%; 5-sec: 32.2%; combined: 35%). Second, very few participants were categorized as having used the *delta rule*. The infrequent assignment of the *delta rule* suggests that participants are modest judgers' of contingency. Third, participants were more likely to use simpler rules to judge contingency when task difficulty was increased. Shaklee and Mims' (1982) increased task by adding an additional counting task, while the current experiment manipulated the event state duration. Nevertheless, the percentage of participants that used the *cell a rule* in the 3-s slide duration condition (i.e., 27.9%) was more than twice that in the 5-s slide duration condition (i.e., 13.6%).

Bivariate Correlations. To test the fourth hypothesis, Pearsons correlation coefficients (r) were calculated on the absolute values of judgments of strength, Δp, and $\Delta p'$. The resulting r values were low (Table A6), which may have resulted from misuse of the judgment of strength scale. Several measures of central tendency suggest that participants misused the judgment of strength scale and awarded noncontingent relations a value of 50 ($M = 49.40$, $Mdn = 50$, Mode = 50.0; note that $\Delta p = 0.0$). Even participants that correctly estimated the frequencies of the 6*a*, 6*b*, 6*c*, 6*d* contingency misused the scale ($n = 44$, $M = 53.50$, $Mdn = 50$, Mode = 50.0). Misuse of the unidirectional scale resulted in weak correlations because of the minimal changes in judgments that resulted from changes in Δp. That is, the mean judgments for the noncontingent problem sets were just as high as those for the positive and negative contingencies (see Table A1).

Conclusions

To summarize, two bodies of evidence suggested that participants differentially make errors when recalling the frequencies of causal information. First, the magnitude of errors for *cell a* was less than those for the remaining cells. The relative magnitude of errors for *cell a* is consistent with its ranking of cell importance (i.e., *cell a* > *cells b*, *c*, and *d*; see Crocker, 1982; Maldonado, Catena, Candid, & Garcia, 1999; Mandel & Lehman, 1998; Wasserman, Dorner, & Kao, 1990; Wasserman & Kao, 1993; White, 2003); however, the magnitude of errors did not increase across all four cells (cf. Shaklee & Mims, 1982). Some evidence did suggests that the magnitude of errors increased across *cells a*, *b*, and *c*; however, only the difference between *cells a* and *b* was significant. Perhaps the difference in beliefs about cell importance is greater between *cells a* and *b* than between *cells b*, *c*, and *d*.

Second, the magnitudes of overestimations were greatest for *cell a* and the magnitudes of underestimations were greatest for the remaining cells (i.e., as indicated by the mean signed cell deviations). The mean signed unit errors also indicated that participants' estimates were more likely to be overestimations for *cell a* and underestimations for all remaining cells. The shared signs in the previously mentioned variables suggest that participants were not only more likely to overestimate *cell a* but the magnitude of those errors were likely to be greater than those for the underestimations. The exact opposite can be said of the latter three cells.

Two factors associated with the use of the unidirectional scale likely contributed to the nonsignificant correlations in the current experiment. First, the unidirectional scale made it more appropriate to use absolute values when calculating the correlations. The absolute values restricted the range of the included values (e.g., with Δp the range was 0 to .62 instead of -.62 to .62). Perhaps if negative values were used the resulting correlations would have been stronger. Second, misuse of the unidirectional scale resulted in judgment insensitivity (discussed in greater detail in the Discussion).

Peterson (1980) also proposed an explanation that might account for the misuse of the unidirectional scale. He argued that erred judgments of noncontingent relations may be due to participants expecting to judge a relation between events, not the lack of a relation. It seems unlikely that Peterson's (1980) explanation accounts for the erroneous judgments in the current experiment because the concept of unrelated events was made explicitly clear, which he found was one way to remove this problem in naive participants. Therefore, it is more likely that participants failed to understand the scale and not that they failed to understand the concept of unrelated events.

The difference in the mean relative difficulty item was not significant. Responses were as expected; however, the difference failed to reach statistical significance. That failure may have been due to the placement of the Comparison Set. Perhaps after having judged 12 problem sets, participants had already adopted a strategy to aid their recall. Therefore, the impact of adding or removing 2 s from each event state may have been attenuated. Perhaps if the Comparison Set was included earlier in experiment, the difference would have been greater in magnitude. That is, before repeated practice estimating cell frequencies permitted efficient strategy use.

The exact strategies that participants may have used to aid recall, if any at all, can only be speculated. Even if qualitative feedback had been obtained, research suggests that humans are extremely biased and are often inaccurate when commenting on their own mental processes (Nisbett & Ross, 1980; Nisbett & Wilson, 1977). Therefore, analyzing the data might serve as a more appropriate means to determine whether strategies were used to aid recall.

A re-analysis of the data suggested that participants used the total number of event states to aid their recall. First, a high percentage of participants' estimates for each problem set totaled 24 (see % Correct Sum in Table A7). Second, the mean sum of participants' estimates (Σfs in Table A7) closely approximated 24 for each of the 12 problem sets. Last, the mean sum of the signed deviations (ΣDeviations in Table A7) closely approximated zero for each problem set. If the sum of the signed deviations equals zero, it implies one of three things: (1) that no errors occurred, (2) that underestimation errors were offset by an equal number of underestimation errors, or (3) vice versa. None of the previously mentioned statistics can be used to imply that any of the cell estimates were correct; however, they do suggest that the majority of participants' estimates were governed by the total number of event states. Therefore, it is uncertain whether the estimates in the current experiment reflect what participants actually stored in memory.

Several concerns raised from the current experiment will motivate modifications made in Experiment 2. First, providing participants with the total number of event states seemed to have promoted strategy use. Therefore, attempts will be made to reduce the likelihood that participants use similar strategies to aid recall. Second, the unidirectional scale proved to be problematic. Therefore, a bidirectional scale will be used with hopes that it improves judgment competency. Last, placement of the Comparison Set will be manipulated to determine whether its impacted the relative difficulty of recalling cell estimates.

CHAPTER II

EXPERIMENT 2: KNOWLEDGE OF PROBLEM SET TOTALS

Participants may have used strategies to aid their recall of event states in Experiment 1. Event states, collectively referred to as causal information, are the four forms of information that define the extent that variables are related. First, participants may not have encoded the occurrences of a particular event state when encoding causal information. All of the problem sets were defined by 24 events. Therefore, a participant could determine the frequency of the event state that was not encoded by subtracting the sum of three encoded estimates from 24. If the sum of the three estimates equals the sum of their actual frequencies, the frequency of the disregarded event can be correctly determined.

Second, participants might have adjusted their cell estimates to ensure that they totaled the correct number of events. Participants made an average of 3.4 errors per problem set [6]. However, more than 80% of participants' estimates totaled 24 events within problem sets. These statistics suggest that participants made an effort to ensure that their estimates totaled 24 events (see Table A7 for more supportive statistics).

[6] Errors refer to the sum of the deviations between the actual and recalled frequencies of causal information.

However, adjusting cell frequencies in this manner does not guarantee an improvement in cell accuracy.

Third, participants might have used their hands to aid memory of causal information. Experimenters frequently observed participants counting on their hands when observing problem sets. One participant shared that she counted the cell frequencies on her four knuckles and immediately entered them into the contingency table. Shaklee and Mims (1982) made no mention of whether participants were permitted to count on their hands or whether they did so. If a participant used his or her hands effectively, the task should more appropriately be viewed as a counting task and not the intended recall task.

The possibility that participants used strategies to aid recall is a concern because they are unlikely to occur in naturalistic settings. Consider a police officer who, after a long day, wishes to assess the likelihood that sports cars are driven by college students. If the officer's assessment is limited to the observations made throughout that day, it is unlikely that he or she would know the total number of cars in the sample. It is also unlikely that the officer would have used his or her hands to count the instances as they occurred earlier in that day.

Two modifications in the current experiment were made in an effort to reduce strategy use. First, participants were told to rest their hands while viewing slides. It is possible that the first manipulation might suggest how to count cell frequencies to participants who might not have normally thought of such a strategy on their own. Therefore, participants were only told to rest their hands and not why they were to rest their hands (i.e., so they would not count cell frequencies on their hands). Second, the current experiment manipulated whether participants are told the total number of event states. To ensure that the second manipulation is maximally effective, the total number of event states varied across problem sets ($M = 23.69$ Events).

The Comparison Set in Experiment 1 was not as successful as hoped. The time to view each event state in the Comparison Set was either longer or shorter than that used for the 12 problem sets. Participants rated the change in relative difficulty of recalling the cell frequencies of the Comparison Set with that of the problem sets. The placement of the Comparison Set was thought to have resulted in the failure to obtain significant differences between conditions. Therefore, the placement of the Comparison Set will be manipulated in the current experiment. That is, participants will either view the Comparison Set early or late in the experimental sequence.

Three hypotheses are proposed for the current experiment. First, it is expected that the mean absolute cell deviations replicate the pattern of differential errors reported by Shaklee and Mims (1982). Second, it is expected that the sum of participants' estimates are more variable in the condition that they are not told the total number of events. Last, it is expected that responses on the relative difficulty scale are more extreme when participants view the Comparison Set early in the experiment.

Method

Participants

Fifty-seven students (14 males and 43 females, mean age = 19.19) enrolled in lower-level psychology courses at the University of New Hampshire participated in the current experiment. All participants received course credit for their participation. Fiftysix percent of the participants were freshman and eight percent were upperclassmen (i.e., 6 and 2%s were juniors and seniors, respectively).

Design

A mixed design was used that consisted of a within-subjects factor (Cell) and two between-subjects factors (Set Placement & Set Information). Cell (4 levels), the within-subjects factor, refers to participants' estimates of the four forms of causal information. The problem sets from Shaklee and Mims' (1982, Experiment 1) experiment were modified so that the cell frequencies were not all defined by 24 events in such a way that they still permitted the rule-analysis technique. Table A8 displays the cell frequencies, total number of events, and Δp for each problem set. Problem Sets 1 through 12 are listed in Table A8 in the order experienced by participants. Problem sets presented slides using a 3-s event state duration. Abstract and contextual problem sets were depicted using the same cues from Experiment 1.

Set Placement (2 levels) refers to whether participants viewed the Comparison Set early or late in the experiment. The Comparison Set, which had a 5-s event state duration, occurred either after the first problem set (i.e., the early condition) or after the 11th problem set (i.e., the late condition). The cues that defined the Comparison Set (Context II in Table A8) depicted the relation between snowing and seeing a bird perched outside a window.

Set Information (2 levels) refers to whether participants were told the total number of event states that defined each problem set. Participants in the known set information condition were told this information, while those in the unknown set information condition were not.

Data were collected in groups that ranged in size from 4 to 15 participants. Three of the four experimental conditions were completed in one session. However, the late-unknown condition was completed in two experimental sessions. As a result, assignments to experimental conditions were determined by the session that participants attended. Table A9 lists the number of participants in each experimental condition.

Measure and Items

Measures were identical to those used in Experiment 1 with several exceptions. First, the short description for the Comparison Set read as follows:

> Some birds flourish when it snows. For example, seeing these birds would most likely depend on whether it was snowing outside. However, some birds migrate in the winter and will not be seen when it snows. These birds would most likely be seen when it is not snowing. Still some other birds are likely to be seen whether or not it is snowing or not. Use these slides to decide whether this bird is one that is most likely seen when it is snowing, not snowing, or if it doesn't matter whether it is snowing.

Second, judgments of contingency were recorded on a measure that obtained the strength, direction, and relative likelihood of the output variable with one response. Judgments of contingency were recorded on a bidirectional scale ranging from -100 (*much less likely*) to +100 (*much more likely*). The following is an example of the judgment of contingency item:

When the plant was watered (as depicted in the slides), it was

|----I----|----I----|----I----|----I----|----I----|----I----|----I----|----I----|----I----|----I----|

-100	-60	-30	±0	+30	+60	+100
much less likely	somewhat less likely	a bit less likely	just as likely	a bit more likely	somewhat more likely	much more likely

to also see plant growth. Draw a line on the scale that best completes the sentence.

Items were modified to reflect the cues depicted in the abstract problem sets and the Comparison Set.

Last, the measures used in the set information conditions differed from one another. The following sentence could be read beneath each contingency table in the known set information condition: *There are a total of x events in this problem set.* Where x is the total number of event states in that problem set. This sentence was not present in the unknown set information condition.

Procedure

The experimenter read the instructions at the start of the session which, Deviations from Experiment 1 began in the second paragraph and read as follows:

> After viewing a set of instances---which may---or may not be about snow and birds--- you will be asked to recall the number of each type of instance. In a blank table---like the one labeled A, B, C and D on the first page---you will be asked to enter the number of times each of the four event types occurred. In other words---you will be shown a problem set---that consist of any combination of the four types of instances just discussed. You will then be asked to recall from memory the number of times each of the four event types were observed---(in other words)---how many times you remember seeing each of the four events---and enter those numbers into the provided blank table.
>
> We ask that during the viewing of slides that you put any writing utensils down and rest your hands while viewing the problem set. I think we all know that memory is not perfect---we wish to learn how memory works---so pay close attention and only refer to your memory---not that of your neighbors.
>
> You will also be asked to judge the relationship between the two events. For this item---we ask that you draw a line on the scale that best describes the relationship between the events in question. A value of positive 100 indicates that the events always occurred together---meaning that every time you looked out the window---it was snowing and a bird was perched outside your window {Show slide of *cell a*}---or it wasn't snowing---and no bird was perched outside your window {Show slide of *cell b*}---this would be a perfect positive relationship. Snowing and seeing a bird always occurred together.
>
> A value of negative 100 indicates that the events never occurred together---meaning that every time you looked out the window---it was either snowing and no bird was perched outside your window {Show slide of *cell c*}---or it wasn't snowing---and a bird was perched outside your window {Show slide of *cell d*}---this would be a perfect negative relationship. Snowing and seeing a bird never occurred together.

> A value of zero indicates that you were equally likely to see a bird perched outside your window---whether it was snowing or not. In other words---snowing had no effect on seeing a perched bird.
>
> As you might have guessed---we will not show you a strict relationships---that would be too easy---the relationships in this experiment---will be somewhere in the middle of the two extremes.
>
> If you have any questions concerning the procedure described thus far, please ask them now. Remember---this is not a competition---you should not talk to---consult with---or observe the work of others. Please do not look back at previously completed problems. You will complete a total of <u>thirteen</u> problems throughout the experiment. And remember---we ask that while viewing the slides that you---put down any writing utensils down and rest your hands.

Hyphens and underlined font were points at which the researcher paused or emphasized statements to facilitate participants' comprehension. All other aspects of the experiment were conducted in the same manner described in Experiment 1.

Results

Cell Deviations

The discrepancies between the actual cell frequencies and participants' estimates were used to create three variables that provide different information about the data: mean absolute cell deviations, mean signed cell deviations, and signed unit errors (See Results in Chapter I for more details). Two sets of paired-samples *t* tests contrasted mean cell deviations, absolute and signed, to determine whether the pattern of cell differences replicated Experiment 1 (i.e., $a \neq b$; $a \neq c$; $a \neq d$; Set 1) or Shaklee and Mims' (1982) experiment (i.e., $a \neq b$; $b \neq c$; $c \neq d$: Set 2). Tests were conducted on the data when collapsed across all conditions and separately for each set information condition.

Mean Absolute Cell Deviations. Prior to conducting the planned comparisons, a multivariate analysis of variance (MANOVA) was conducted on the mean absolute cell deviations. The independent variables (i.e., set placement and set information) served as the MANOVA's factors. Set Information, Set Placement, and their interaction had no effect on any of the four cells (all *p*s $\geq$ .34).

Mean absolute cell deviations are displayed in the top panel of Table A10. When data were collapsed across conditions, the first set of tests revealed that the mean absolute cell deviation for *cell a* was not different than those obtained for *cells b* or *d, cell b, t*(56) = -1.12, M_{SE} = 0.07. *p* = .27; *cell d, t*(56) = -0.86, M_{SE} = 0.08, *p* = .39. However, the mean absolute cell deviation for *cell a* was significantly less than that obtained for *cell c*, *t*(56) = -2.49, M_{SE} = 0.09, $p < .05$. The second set of tests indicated that the mean absolute cell deviation for *cell b* was not different than that obtained for *cell c*, *t*(56) = -1.73, M_{SE} = 0.09, *p* = .09. However, the mean absolute cell deviation for *cell c* was significantly different than that obtained for *cell d*, *t*(56) = 2.00, M_{SE} = 0.08, *p* = .05.

The patterns of significant differences were different from one another in the set information conditions. In the known set information condition, the pattern was similar to that reported when data were collapsed across all conditions. That is, the first set of tests indicated that the mean absolute cell deviation for *cell a* was not different than those obtained for *cells b* and *d*, *cell b, t*(28) = -1.74, M_{SE} = 0.09, *p* = .092; *cell d*, *t*(28) = -0.98, M_{SE} = 0.10, *p* = .34. The mean absolute cell deviation for *cell a* was, again, significantly different than that obtained for *cell c*, *t*(56) = -2.57, M_{SE} = 0.14, $p < .05$. The second set of tests indicated that the mean absolute cell deviation for *cell b* was not different than that obtained for *cell c*, *t*(56) = -1.60, M_{SE} = 0.119, *p* = .121. However, the mean absolute cell deviations for *cell c* was significantly greater than that obtained for *cell d*, *t*(56) = 2.44, M_{SE} = 0.101, $p < .05$. In the unknown set information condition, none of the tests indicated that cells were significantly different from one another (all *p*s $\geq$ .39).

Mean Signed Cell Deviations. A MANOVA was also calculated on the mean signed cell deviations, which again served as a manipulation check. Set Information had a significant main effect on *cell b*, $F(1) = 5.63$, $M_{SE} = 2.14$, $p = .02$. Set Placement also had a significant main effect on *cell b*, $F(1) = 7.08$, $M_{SE} = 2.69$, $p = .01$.

Mean signed cell deviations are displayed in the bottom panel of Table A10. When data were collapsed across all conditions, the first set of test indicated that the mean signed cell deviation for *cell a* was significantly different than those obtained for the remaining three cells, *cell b*, $t(56) = 3.86$, $M_{SE} = 0.13$; *cell c*, $t(56) = 6.42$, $M_{SE} = 0.12$; *cell d*, $t(56) = 5.15$, $M_{SE} = 0.14$; all *p*s < .05. The second set of tests revealed that the mean signed cell deviation for *cell b* was significantly different than that obtained for *cell c*, $t(56) = 2.36$, $M_{SE} = 0.12$, $p < .05$. However, the mean signed cell deviations for *cell c* was not different than that obtained for *cell d*, $t(56) = -.58$, $M_{SE} = 0.13$, $p = .56$.

In the known set information condition, the first set of tests revealed that the mean signed cell deviation for *cell a* was not significantly different than that obtained for *cell b*, $t(28) = 1.67$, $M_{SE} = 0.16$, $p = .11$. However, the mean signed cell deviation for *cell a* was significantly different than those obtained for *cells c* and *d*, *cell c*, $t(28) = 4.05$, $M_{SE} = 0.17$; *cell d*, $t(28) = 4.07$, $M_{SE} = 0.167$, all *p*s < .05. The second set of tests revealed that the mean signed cell deviation for *cell b* was significantly different than that obtained for *cell c*, $t(28) = 2.81$, $M_{SE} = 0.154$, $p < .05$. However, the mean signed cell deviation for *cell c* was not different than that obtained for *cell d*, $t(28) = -.12$, $M_{SE} = 0.171$, $p = .90$.

In the unknown set information condition, the first set of tests revealed that the mean signed cell deviation for *cell a* was significantly different than that obtained for each of the remaining three cells, *cell b*, $t(27) = 3.73$, $M_{SE} = 0.21$; *cell c*, $t(27) = 5.00$, M_{SE}

= 0.18; *cell d*, $t(27) = 3.31$, $M_{SE} = 0.22$; all *ps* < .05. The second set of tests revealed that no significant differences existed between *cells b* and *c* or between *cells c* and *d*, *cells b* and *c*, $t(27) = .63$, $M_{SE} = 0.17$; *cells c* and *d*, $t(27) = -.66$, $M_{SE} = 0.21$,all *ps* ≥ .52.

Signed Unit Errors. Table A11 displays for each cell the percentage of estimates that were over-, under-, or correct estimations. An independent-samples *t* test indicated that the percentages of correct estimations in the set information conditions were not significantly different from one another, $t(6) = 0.54$, $Diff_{SE} = 2.47$, $p = .33$.

Table A11 also displays the mean signed unit errors, which can range from -1.0 to +1.0. Positive values indicate that errors were more likely to be overestimations, while negative values indicate that they were more likely to be underestimations. No hypotheses or planned comparisons were proposed for the mean signed unit errors. Therefore, the data will be reviewed in the Discussion.

Sum of Estimates

The mean sum of participants' estimates was calculated within problem sets, hereafter referred to as the sum of estimates (labeled as Actual in Table A12). Table A12 lists the minimum and maximum of the sum of estimates (labeled as Min and Max in Table A12, respectively). The minimum and maximum for each problem set were used to calculate its range, hereafter referred to as the range of the sums. An independent samples *t* test indicated that the range of the sums was significantly different from one another in the set information conditions, $t(24) = -5.85$, $Diff_{SE} = 1.63$, $p < .05$.

The percentage of participants whose sum of estimates totaled the correct number of event states are also listed in Table A12 (labeled as % Correct Sum). A second *t* test indicated that the percentage of participants whose sum of estimates totaled the correct number of events were significantly different from one another in the set information conditions, $t(24) = 19.07$, $Diff_{SE} = 3.10$, $p < .05$.

Relative Difficulty in Recall

Responses on the relative difficulty item were used to determine whether differences existed in the set placement conditions. An independent-samples *t* test indicated that the values in the set placement conditions were not significantly different from one another, $t(41) = 1.73$, $Diff_{SE} = .28$, $p = .09$.

Judgments

Multivariate Analysis of Variance. A third MANOVA, again serving as a manipulation check, determined whether the independent variables had an effect on the judgments of contingency. Only Set Placement indicated that it had a significant main effect on the judgments of contingency for Problem Set 12, $F(1) = 4.57$, $M_{SE} = 7140.90$, $p < .05$. No other main effects were significant.

Rule-Analysis Technique. Judgment accuracy for the rule-analysis technique was determined in a slightly different manner. Participants were said to have successfully judged a problem set if they correctly determined its direction. However, the direction of participants' judgments was determined by the judgment of contingency item instead of relative likelihood item (cf. Experiment 1). All other aspects of the rule-analysis technique were performed as described in Experiment 1.

Bivariate Correlations. Table A14 lists the resulting *r* values of the correlations between judgments of contingency, Δp, and $\Delta p'$. When collapsed across all conditions, judgments of contingency were significantly correlated with Δp, $r = 0.23$, $n = 657$, $p < .05$. Judgments of contingency were also significantly correlated with $\Delta p'$, $r = 0.29$, $n =$

652, $p < .05$. The difference between the two previously mentioned correlations approached, but failed to reach, conventional significance, $t(676) = -1.35$, $p = .08$ [one-tailed test]. Last, Δp and $\Delta p'$ were also significantly correlated with one another, $r = 0.82$, $n = 679$, $p < .05$.

In the known set information condition, a similar pattern of results was found. Judgments of contingency were significantly correlated with Δp, $r = 0.21$, $n = 341$, $p < .05$. Judgments of contingency were also significantly correlated with $\Delta p'$, $r = 0.27$, $n = 339$, $p < .05$. However, the difference between the two previously mentioned correlations was significant, $t(343) = -1.76$, $p < .05$ [one-tailed test]. Last, Δp and $\Delta p'$ were also significantly correlated with one another, $r = 0.81$, $n = 346$, $p < .05$.

In the unknown set information condition, a similar pattern of results was also found. Judgments of contingency were again significantly correlated with Δp and with $\Delta p'$, Δp, $r = 0.24$, $n = 316$; $\Delta p'$, $r = 0.31$, $n = 313$, all *p*s $< .05$. The difference between the two previously mentioned correlations was also significant, $t(333) = -2.42$, $p < .05$ [one-tailed test]. In addition, Δp and $\Delta p'$ were also significantly correlated with one another, $r = 0.83$, $n = 333$, $p < .05$.

Discussion

Cell Deviations

Mean Absolute Cell Deviations. Two sets of paired-samples *t* tests tested the first hypothesis. The two sets of tests determined whether the data supported the results of Experiment 1 or Shaklee and Mims' (1982) experiment. For the most part, there were no significant differences in the magnitudes of errors between cells in the current experiment. Two exceptions being that the magnitude of errors for *cell c* was significantly higher than those obtained for *cells a* and *d* in all but the unknown set information condition. Like Experiment 1, the magnitude of errors increased across the first three cells of the contingency table (i.e., *cells a*, *b*, and *c*).

Mean Signed Cell Deviations. Mean signed cell deviations for each cell were examined to determine whether the magnitude of overestimations was greater than that for underestimations, or vice versa (see bottom panel of Table A10). Like Experiment 1, participants consistently overestimated the frequency of *cell a* to a greater extent than they underestimated its frequency and the exact opposite was true for the remaining three cells. Differences in the mean signed cell deviations between *cell a* and each of the remaining three cells were significant. However, there was one exception to the mentioned trend. The magnitudes of errors for *cell b* were greatest for overestimations in the known set information condition and, as a result, the difference between *cells a* and *b* was not significant.

The MANOVA revealed two significant main effects. First, the mean signed cell deviation for *cell b* was higher in the unknown set information condition ($M = .15$, $SD = .55$) than in the known condition ($M = -.21$, $SD = .73$). Second, the mean signed cell deviation for *cell b* was higher in the late set placement condition ($M = .18$, $SD = .70$) than in the early condition ($M = -.24$, $SD = .57$). Further analysis indicated the late-known condition was the only one in which the mean signed cell deviations were positive ($M = .41$, $SD = .57$). The mean signed cell deviations in the remaining three conditions (i.e., late-unknown, early-known, early-unknown) were all negative (*M*s = -.04, -.09, & -.42, respectively). It is not apparent why the mean signed cell deviations were positive for only one experimental condition. However, it does suggest that the tendency for the magnitude of errors for *cells b*, *c*, and *d* to be greatest for underestimations is a robust effect.

Signed Unit Errors. The signed unit errors were used to determine if any differences existed in the likelihood of correctly estimating cell frequencies in the set information conditions. There was no difference in the percentages of correct estimations between the two set information conditions, which suggests that knowledge of the total number of event states was not used in a way to improve cell accuracy. Ensuring that cell estimates totaled the correct number of events does not guarantee an increase in cell accuracy. The data from the current experiment are consistent with that notion.

The mean signed unit errors were also examined because its sign can be used as an index of the tendency to over- or underestimate cell frequencies. Participants were more likely to overestimate confirming cases and underestimate disconfirming cases (Table A11). The one exception occurred for *cell b* in the known set information condition ($M = .03$, $SD = .77$). The signs of the mean signed unit errors and the mean signed cell deviations were consistent with one another for *cells a*, *b*, and *c*. However, the mean signed unit errors were positive for *cell d* and negative for the mean signed cell deviations. Their differences indicate that although participants were more likely to overestimate the frequency of *cell d* the underestimations were of greater magnitude. This difference may at first seem problematic; however, the present investigation is the first to report mean signed unit errors. As a result, it is not known whether any consistent trends are to be expected with this variable.

Sum of Estimates

To test the second hypothesis, independent-samples *t* tests were performed, across set information conditions, the range of the sums and the percentages of participants whose sum of estimates totaled the correct number of event states (see Table A12). First, the range of the sums was significantly larger in the unknown set information condition (M = 14.00, SD = 4.42) than in the known set information condition (M = 4.46, SD = 3.89). Second, the percentage of participants' whose sum of estimates totaled the correct number of events was significantly higher in the known set information condition (M = 88.88, SD = 5.50) than in the unknown set information condition (M = 29.94, SD = 9.74). Collectively, the *t* tests suggest that the sum of estimates were more variable when participants did not know the total number of events.

Evidence that participants adjusted the frequencies of their estimates was of interest in the current experiment. Other than significant differences in the likelihood that estimates totaled the correct number of events, no other differences were found between the set information conditions. There was no difference in the percentage of correct estimations and both set information conditions produced the same trend across cells in the mean absolute and the mean signed cell deviations. Therefore, it is unlikely that the differential patterns of errors reported in Experiment 1 were due to participants adjusting cell frequencies.

Relative Difficulty of Recall

To assess the third hypothesis, participants' responses on the relative difficulty scale were examined. The difference between the early (M = .78, SD = .85) and late (M = .30, SD = .98) conditions was in the predicted direction. However, the difference failed to reach statistical significance. The data suggest that by the time participants viewed the Comparison Set late in the experiment they had learned to efficiently use a strategy to aid their recall. The data also suggest that early in the experiment participants either have not adopted a strategy yet or have not learned to use it efficiently enough to decrease the perceived difficulty of the recall task.

The lack of a significant difference between the set placement conditions on the relative difficulty item may have been due to the size of the scale. The relative difficulty item, a 5-point Likert scale, may have been too small to obtain a significant difference across conditions; that is, especially considering that values were expected on the positive end of the scale in both conditions. Had one condition rated an increase in perceived task difficulty and the other condition a decrease, as in Experiment 1, perhaps a significant difference would have been obtained. In retrospect, had the Comparison Sets been implemented earlier in Experiment 1, as in the current experiment, perhaps a significant difference would have been obtained then.

Judgments

Rule-Analysis Technique. The percentage of participants categorized as having used each rule is in many ways similar to that reported by Shaklee and Mims (1982; Experiment) and Experiment 1 (see Table A13). First, approximately one third of the participants were deemed as *unclassifiable*. Second, participants were more likely to use simpler rules to judge contingency. Third, the percentages of participants categorized as having used the *delta rule* were rare. However, the majority of participants in the current experiment were categorized as having used the *a-versus-b rule* and not the *cell-a rule* (cf. Experiment 1).

Bivariate Correlations. Pearson correlation coefficients were calculated on judgments of contingency, Δp, and $\Delta p'$ to test the last hypothesis. The differences in the correlations were in the predicted directions. Indicating that the correlations between judgments and $\Delta p'$ were significantly stronger than those between judgments and Δp. However, the difference in correlations did not reach statistical significance when the data were collapsed across all conditions. This effect will be discussed in greater detail in the Conclusions.

Conclusions

To summarize, the failure to obtain significant differences in the mean absolute cell deviations suggests that the differential pattern of errors is a weak effect. However, the failure to replicate the effect with significant differences is reminiscent of what occurred in Shaklee and Mims' (1982) experiments. Significant differences between cells were not consistently obtained in their two experiments. In fact, only two comparisons were reliably significant across experiments. First, the magnitude of errors for *cell a* was less than that obtained for *cell b*. Second, the magnitudes of errors for the cells in the top row were less than those in the bottom row. However, a monotonic increase in the magnitude of errors was found in both experiments. Therefore, the failure to support the first hypothesis through significant differences might not be of great concern.

It was argued in Experiment 1 that the same of pattern of results in two of the three reported analyses was an indication of the robustness of the effect (i.e., when data were collapsed across conditions and in the 3-s slide duration condition). Similar results were also found in the current experiment (i.e., when the data were collapsed across conditions and in the known set information condition) and the consensus is now that the differences in the magnitude of errors is a weak effect. Therefore, it is more likely that the effect in one of the experimental conditions was responsible for the shared trend when data were collapsed across conditions.

The fact that participants were more likely to overestimate confirming cases and to underestimate disconfirming cases may be consistent with that reported by Crocker and Taylor (1978; as cited by Alloy & Tabachnik, 1984). As suggested by Peterson (1980), participants arrive at experiments with strong expectations that they will judge a relationship between the depicted events. It might also be that participants expect to judge a positive relationship, especially when contextual cues are used that depict real world relations. Crocker and Taylor (1978; as cited by Alloy & Tabachnik, 1984) reported that participants consistently overestimated the frequencies of information that were consistent with their expectations and underestimated information that were inconsistent with their expectations. If participants expected to judge a positive relation, then the confirming cases are the expectation consistent information and the disconfirming cases are the expectation inconsistent information. A re-analysis of the data indicated that participants overestimated confirming cases and underestimated disconfirming cases for contextually defined problem sets (*M*s = .11, -.07, -.07, and .07 for cells a, b, c, and d). The means indicated a different pattern across cells when the problem sets were defined by abstract cues (*M*s = .17, .06, -.03, and -.04 for cells a, b, c, and d). However, the question remains as to why a similar tendency was not reported in Experiment 1 for the participants that viewed 12 contextually defined problem sets.

As predicted, the correlations between contingency judgments and $\Delta p'$ were significantly higher than those between judgments and Δp. This effect suggests that judgments are more consistent with that stored in memory than with what actually occurred. This effect is the biggest contribution of the current investigation. Therefore, an extended discussion of its results and implications will be conducted.

The resulting *r* values might have been artifacts of pooling data across the set information conditions. Further analyses, reported here for the first time, revealed that the correlations between judgments and $\Delta p'$ (i.e., *r*s for $\Delta p'$ equaled .36, .52, .44, & .52) were stronger than those between judgments and Δp in each experimental condition (i.e., *r*s for Δp equaled .27, .45, .35, & .42 in the early-known, late-known, early-unknown, and late-unknown conditions, respectively). Each experimental condition was a separately conducted experimental session; therefore, each session can be viewed as independent replication (Baron & Perone, 1998; Sidman, 1988).

The *r* values obtained in the current experiment were relatively weak in comparison to previous research (cf., Allan & Jenkins, 1980; 1983; Wasserman et al., 1993; Wasserman, Dorner, & Kao, 1990). One might argue that the weak *r* values are an indication of poor experimental control; however, a closer look at the data suggests that this not the case. Problem sets were borrowed from Shaklee and Mims' (1982; see also Shaklee & Tucker, 1980) experiment and were designed for the rule-analysis technique, which exploits the circumstances in which the rule-based strategies produce inaccurate judgments. Problem sets designed in such a manner may be particularly difficult to judge. If the problem sets are difficult due to their design, then the correlations between judgments and Δp should become progressively weaker as the subset's sophistication increases.

Table A15 lists the resulting *r* values between judgments of contingency, Δp, and $\Delta p'$ for the four subsets of the rule-analysis technique in the order of increasing sophistication. The *r* values between judgments and Δp for the *cell a* and *a-versus-b subsets* are consistent with that reported in previous research (cf. Allan & Jenkins, 1980; 1983). However, the correlations between judgments and Δp progressively decreased and became negative as sophistication increased. The negative *r* value in the Δp *subset* is consistent with what one would expect because the rule-analysis technique defined judgment accuracy by the direction of participants' judgments (Shaklee & Mims, 1982; Shaklee, 1983). Therefore, the overall weak correlations obtained in the current experiment can be accounted for by the use of the problem sets from the rule-analysis technique.

The results of the current experiment are most consistent with a rule-based account of contingency judgment. The associative model of contingency judgment assumes that memory of causal information is irrelevant for judging contingency (Baker, Murphy, & Vallee-Tourangeau, 1996) and that judgments are formed through an experiential mechanism. It is unclear how the associative model can account for a tendency for judgments to be more consistent with memory of events than with their actual frequencies. According to the associative model, there is no reason for the relationship between judgments and $\Delta p'$ to be stronger than that between judgments and Δp.

The results of the current experiment also suggest that rule-based strategies, or at least an approximation of them, are used to judge contingency. It is unlikely that a strict adherence to any rule is used by participants to judge contingency. However, it is likely that individuals use a set of questions that, upon being answered, permits an approximation of that which is predicted by the rule-based strategies. The fact that judgments are more highly correlated with a rule based on one's memory, which is an approximation in its self, suggests that rough estimates of rule-based strategies are used to judge contingency.

The results of the current experiment might also suggest that $\Delta p'$ is a more suitable index from which to assess the accuracy of participants' judgments. Many researchers have disputed whether Δp, or any normative statistic, is an appropriate index from which to assess the accuracy of judgments (Mandel & Lehman, 1998; Shaklee and Mims, 1983). Although the rule by which $\Delta p'$ is calculated is the same as Δp, use of participants' estimates as its inputs would arguably change its status as a normative statistic (see Mandel & Lehman, 1998 for a discussion of normative ness). It also seems more reasonable to assess judgment accuracy on what the individual believes they are judging as opposed to what they are judging. The only drawback to using $\Delta p'$ is that obtaining participants' estimates is very cumbersome. However, use of $\Delta p'$ might be even more beneficial in situations in which larger discrepancies exists between memory and the relevant information or when judgments are not based on information provided by the experimenter.

Discrepancies between memory and actual information might be relevant to illusory correlations. Illusory correlation has been defined as the tendency to perceive a relation between events that are not related or are related to a lesser extent than that which it is judged (Chapman, 1967). Illusory correlations might more appropriately be viewed as judgments based on a mental representation that is quite discrepant from that which exists. An inaccurate mental representation of events might be the mechanism by which illusory correlations are formed. If illusory correlations are mediated by inaccurate mental representations, then its definition should be broadened. The definition of an illusory correlation must be broadened because an inaccurate mental representation can result in an increase or decrease in the magnitude of the perceived relation.

Hamilton and Gifford (1976; Experiment 1) has similarly argued that illusory correlations are based on inaccurate mental representations. Their experiment investigated stereotypic person perception, which they argue is the social equivalent to illusory correlations. Participants rated the likeability of members from two groups that differed in their relative size, after members of each group were paired with desirable and undesirable traits. Favorable ratings were biased towards the larger group despite the proportions of desirable and undesirable traits being equal in the two groups. It was concluded that because the pairings of undesirable traits with members of the smaller group were infrequent that their occurrences were a salient event. The salience of the events resulted in its frequency being overestimated, which biased favorable ratings towards the larger group. This finding suggests that perhaps illusory correlations, at least their social equivalents, are accompanied with a similarly biased mental representation of the relevant events.

Summary of Phase I

The first phase of the current investigation collectively addressed five hypotheses. First, cell estimates were examined to determine whether participants differentially made errors when recalling their frequencies. Second, event state duration was manipulated to determine whether it impacted the magnitude of errors. Third, the impact of the event state duration was examined on participants' perceived difficulty when recalling cell frequencies. Fourth, the impact of participants' knowledge of the total number of event states was examined on their cell estimates. Last, Δp and $\Delta p'$ were correlated with contingency judgments to determine their relative dependence on memory of causal information. The following segment summarizes the main results and conclusions relevant to these five hypotheses.

Differential Pattern of Errors

Three variables were used to assess differences across cells. First, the mean absolute cell deviations were used as an index of the magnitude of errors. Shaklee and Mims' (1982), using the same variable indicated that the magnitude of errors increased across *cells a*, *b*, *c*, and *d* in Experiments 1 and 2. The magnitude of errors were only found to increase across *cells a*, *b*, and *c* in the current experiment. The magnitude of errors for *cell d* never once exceeded that for *cell c*. It is believed that the differential pattern in the magnitude of errors is not a robust effect.

Second, the mean signed cell deviations were used as an index of whether the magnitude of errors were greater for under- or overestimations. The current investigation is the first that used the mean signed cell deviations to characterize cell estimates and, as a result, no specific hypotheses were proposed. However, this variable did produce a reliable pattern across Experiments 1 and 2. The magnitude of errors were greatest for *cell a*, while the exact opposite was true for *cells b*, *c*, and *d*. It is believed that *cell a* is a more salient cue, which increases the likelihood that its frequency is overestimated (see Hamilton & Gifford, 1976 for a similar argument).

Last, signed unit errors were used as index of the tendency to over- or underestimate cell frequencies. Originally, the signs of the mean signed unit errors and means signed cell deviations were consistent with one another (Experiment 1). However, in Experiment 2 their signs were different for *cell d.* It is not clear, at this point, whether either pattern is an indication of a real tendency to under- or overestimate cell frequencies. The similarities across experiments for *cells a*, *b*, and *c* do suggest that the tendencies to under- and overestimate cells frequencies are systematic (cf. Smedslund, 1968).

The Effect of Slide Duration

In Experiment 1, the event state duration was manipulated. I concluded that the shorter duration interfered with participants' use of strategies to aid recall. The magnitudes of errors were significantly greater in the condition with the shorter event state duration. It is believed that the longer event state duration permitted more efficient use of strategies, which resulted in there being no differences in the magnitude of errors across cells.

Relative Difficulty of Recall

Participants viewed 12 problem set using the same event state duration; however, a Comparison Set was also viewed that used a different event state duration. Participants indicated whether an increase or decrease in the time to view each event state impacted the perceived difficulty of recalling cell frequencies. In Experiment 1, responses indicated that more time resulted in a relatively easier task (i.e., 5 s vs 3 s) and that less time resulted in a relatively harder task (i.e., 3 s vs 5 s). Only one event state duration was used in Experiment 2. However, participants indicated that additional time was more beneficial when provided earlier in the experiment. It is believed that the limited range of the relative difficulty item, a 5-point Likert scale, contributed to the failure to support this hypothesis with significant differences.

Sum of Estimates

Experiment 2 further addressed whether participants used strategies to aid their recall. Evidence suggested that participants made an effort to ensure that their estimates totaled the correct number of events (Experiment 1). Therefore, whether participants had such knowledge was manipulated in Experiment 2. The sum of participants' estimates provided strong evidence suggesting that, when available, participants used the total number of event states to ensure that their estimates summed to that number. However, the data did not suggest that knowledge of the total number of events impacted the accuracy of participants' cell estimates. The differential patterns of errors were similar whether participants had knowledge of the total number of events, which suggests that the patterns are not due to this particular strategy.

Judgment and Memory

Finally, $\Delta p'$ was used to determine whether its use could compensate for errors in contingency judgments. Participants' misuse of the unidirectional judgment scale Experiment 1 did not permit an assessment of this hypothesis. The bidirectional judgment scale was appropriately used in Experiment 2, which permitted an assessment of the stated hypothesis. Judgments and $\Delta p'$ were more highly correlated with one another than judgments and Δp, which supported the notion that judgments are more highly correlated with memory of the contingency (i.e., $\Delta p'$) than with the actual contingency (i.e., Δp). This finding was extensively discussed.

CHAPTER III

EXPERIMENT 3: KNOWLEDGE OF JUDGMENT TASK

The relationship between memory and contingency judgments might be further understood if a judgment task distinction is acknowledged. Hastie and colleagues (Hastie & Park, 1986; Hastie & Pennington, 1989) argue that there are at least two types of judgment tasks, which differ in their dependence on memory. The two types of judgment tasks, on-line and memory-based, also differ in how people form their judgments. On-line judgment tasks occur when participants are aware of the task at hand and are able to form their judgment as the relevant information is encountered. On the other hand, memory-based judgment tasks occur when participants are made aware of the task after having previously viewed the relevant information and, as a result, are forced to form their judgment on that which can be recalled. Memory and judgments are related on memory-based judgment task, while they are not as highly related on on-line judgment tasks[7].

The work of Hastie and colleagues (Hastie & Park, 1986; Hastie & Pennington, 1989) deals with social judgments. Social judgments refer to inferences made about other people that include, but are not limited to, the following: judging the likely causes of another's behavior (e.g., Hastie, 1984), rating aspects of another's personality (e.g.,

[7] Hastie and colleagues' (Hastie & Park, 1986; Hastie & Pennington, 1989) task distinction is only relevant to contingency judgment experiments that utilized a trial-by-trial procedures, which presents information serially. Experiments that utilize a summarized procedure are not relevant to the current discussion because in those situations judgments are based on the readily available information, not that which was experienced and/or stored in memory.

Hastie & Kumar, 1979), and evaluating another's suitability for a given task (e.g., Hastie & Park, 1986). The social judgment and contingency judgment literatures differ in a number of ways; their most important difference, at least for this investigation, is how judgment accuracy is traditionally assessed. The contingency judgment literature typically quantifies the extent that two variables are related through the use of a statistic (Δp is the most frequently used statistic; e.g., Jenkins & Ward, 1965; Wasserman & Shaklee, 1984). As a result, contingency judgments accuracy has frequently been assessed considering its approximation to the statistic's value (e.g., Allan & Jenkins, 1980; 1983; Alloy & Abramson, 1979; Erlick & Mills, 1967; cf. Shaklee, 1983). On the other hand, social judgments are subjective impressions which often cannot be objectively determined as correct, incorrect, or expressed by a precise statistic. Therefore, it is not known whether the judgment task distinction will be supported by differences in contingency judgments. If the judgment task distinction does apply to contingency judgments, it might account for the failure to obtain a relationship between memory and judgment in previous research (Jenkins & Ward, 1965; Wasserman, et al., 1993).

The contingency judgment literature has almost exclusively examined on-line judgments and no experiment has compared memory-based and on-line contingency judgments; several factors might explain such absences. First, several experiments have failed to find a relationship between memory and contingency judgments (e.g., Jenkins & Ward, 1965; Wasserman, et al., 1993). However, the failed experiments contained on-line judgment tasks in which no relationship is expected. Second, the Associative Model (Shanks, 1985; Shanks & Dickinson, 1987) and judgment task distinction (Hastie & Park, 1986; Hastie & Pennington, 1989) were introduced around the same time. The

Associative Model, which has dominated modern thinking about contingency judgment, assumes that memory of causal information is not necessary for judging contingency (Baker, Murphy, & Vallee-Tourangeau, 1996) and may have halted the limited number of investigations dedicated to the role of memory for judging contingency. Third, conventions in the contingency judgment literature might be partly responsible for the almost exclusive use of on-line judgment tasks. For example, the experimental procedure and judgment scale are often explained to participants prior to viewing the first problem set. Explaining the judgment scale or task would inform participants of the later task and, by definition, result in an on-line judgment. Fourth, experimenters became reluctant to state that participants mentally compute rule-based strategies to judge contingency (see Wasserman, 1990; and cf. Wasserman, 1993). Investigating the relationship between memory and contingency judgments is, in a way, accepting the notion that participants attempt to mentally compute judgments via rule-based strategies.

To the best of my knowledge, only one experiment claims to have investigated memory-based judgments of contingency. In that experiment, Arkes and Rothbart (1985; Experiment 1) presented names on slides that varied according to two dimensions: (1) gender (i.e., male or female name) and (2) relative placement in the alphabet (i.e., begins with a letter from the first [letters a-m] or second [letters n-z] halves of the alphabet). Labeling the rows and columns of a contingency table (e.g., row 1: male names, row 2: female names, column 1: first half of alphabet, column 2: second half of alphabet) permits any name to be appropriately categorized by the two dimensions. Names were selected so that no relation existed between the two dimensions. Participants estimated the frequencies of either a specific row or column of the contingency table (e.g., *recall* *the number of female names*). Judgments were biased by the information that participants were asked to recall (i.e., the specific row or column).

Several aspects of Arkes and Hothbart's (1985) experiment raise doubts as to whether participants actually judged contingency. First, only the frequencies of two cells were recalled. Contingency requires the frequencies of all four cells; one might argue that to obtain a true contingency judgment that the entire summary table should have been estimated. Second, judgments were not predicted by the use of any known rule-based or associative account of contingency judgment. Instead, judgments were predicted by simple inferences made about other cells (see design of Arkes & Rothbart, 1985; Experiment 1). Last, judgments were not of contingency but of the relative likelihood of events (i.e., *who are more likely to have names beginning with a letter in the second half of the alphabet, males or females?*). Judgments of contingency require an estimation of the extent that variables are related, while judgments of relative likelihood ask which of two variables is most likely (see Waldmann & Holyoak, 1997 for a similar argument). These criticisms are intended to express that memory-based contingency judgments have yet to be adequately addressed, not to challenge the results of the experiment.

The current experiment will compare judgments of contingency formed in memory-based and on-line judgment tasks. Shaklee and Mims' (1982) procedure will again be used. However, the points at which participants are informed of the judgment task will differ across the three experimental conditions. Participants in the on-line condition will be informed of the judgment task before viewing the problem set, as typically done. Participants in the two memory-based conditions (i.e., memory-based-PE and memory-based-PR) will both be informed of the judgment task after viewing the problem set. Participants in the memory-based-PE (Post Encode) condition will be informed after the viewing problem set (i.e., after encoding cell frequencies), while those in the memory-based-PR (Post Recall) condition will be informed after completing the summary table (i.e., after encoding and recalling cell frequencies).

Previous research has demonstrated that the magnitude of errors increases across *cells a*, *b*, and *c* when recalling cell frequencies (Experiments 1 and 2; Shaklee & Mims, 1982). The differential pattern in the first three cells is consistent with the notion that subjective cell importance influences the magnitude of errors. Subjective cell importance refers to beliefs about the relevance of each cell for judging contingency (e.g., Wasserman, Dorner, & Kao, 1990). Therefore, it might be that subjective cell importance influences either the recall or encoding of causal information in a way that the accuracy of the more important cells are higher than the lesser important cells. The current experiment may prove useful in determining the source of the differential pattern in the magnitude of cell errors.

It may be that subjective cell importance influences the amount of resources one dedicates to the encoding of causal information. If the differential pattern is due to biased encoding it should only occur in the on-line condition. No differential pattern should occur in either memory-based condition because beliefs about cell importance would not have the opportunity to bias encoding; that is, participants are not aware of the judgment task until after the encoding of causal information.

It may be that subjective cell importance influences the retrieval of causal information (Crocker, 1982). If the differential pattern is due to biased encoding it should not occur in the memory-based-PR condition. However, the pattern should emerge in remaining two conditions. No differential pattern should occur in the memory-based-PR condition because subjective cell importance will not have the opportunity to bias retrieval; that is, participants are not aware of the judgment task until after the encoding of causal information.

The results of Experiments 1 and 2 suggest that the type of scale used to record judgments might influence judgment competency. The current experiment will use multiple measures to assess judgments of contingency. Participants will judge the relative likelihood of the output variable given the presence of the input variable, which, in essence, asks participants whether the presence of the output variable is more or less likely when the input variable is also present. Participants will also judge the extent that the two events are related, which will consist of a scale ranging from -100 to +100 that request an estimation of both the strength and direction of the relation. Although both judgments are of the same contingency, previous research suggests that participants might interpret the questions differently and, as a result, answer them with differing degrees of accuracy (Shaklee & Tucker, 1980). Shaklee and Tucker (1980) used between-groups comparisons, while the current experiment will use within-subjects comparisons to determine judgment scales impact judgment competency.

Several hypotheses are proposed in the current experiment. First, it is expected that the magnitude of errors will increase across *cells a*, *b*, and *c* and replicate previous research (Experiments 1 and 2; Shaklee & Mims, 1982). No hypotheses are proposed about the presence of the differential pattern in the judgment task conditions; however, differences between conditions, as discussed above, might provide insights about its likely source. Second, it is expected that the mean signed cell deviations will be positive for *cell a* and negative for *cells b* and *c*. Third, it is expected that participants' responses on the two judgment scales will differ in judgment accuracy.

Method

Participants

Seventy-six students (26 males and 50 females, mean age = 19.41) enrolled in lower-level psychology courses at the University of New Hampshire participated in the current experiment. All participants received course credit for their participation. More than 90% of the participants were freshman and sophomores (i.e., 41 and 51%s, respectively). Eight percent were upperclassmen (i.e., 6.6 and 1.3 %s were juniors and seniors, respectively).

Design

A mixed design was used that consisted of a within-subjects factor (Cell) and two between-subjects factors (Event Sequence and Type of Judgment). All slides in the current experiment were presented using a 5 s event state duration.

Cell (4 levels), the within-subjects factor, refers to participants' estimates of the four forms of causal information. Participants' estimates were subtracted from the actual frequencies to determine the absolute cell deviations, and signed cell deviations, and signed unit errors for each cell. Participants judged one problem set and its frequencies of the causal information totaled 24 events and were as follows: 10*a*, 4*b*, 2*c*, 8*d*.

Event Sequence (2 levels) refers to the orders that event states were presented: forward and backward. The order of the forward condition was determined by a quasi-random process[8]; its resulting order was as follows: *cadaadbdbadbadadacbadada*. The quasi-random order was reversed for the backward condition.

Type of Judgment (3 levels) refers to the point at which participants were made aware of the judgment task. Participants in each condition (1) viewed the event states, (2) estimated the frequencies of causal information, and (3) judged the portrayed contingency. The point at which the judgment task instructions were read, which informed participants of the judgment task, varied across the three conditions. In the on-line condition, the judgment task instructions were read at the start of the experiment. In the memory based-PE condition, the judgment task instructions were read after viewing the problem set. In the memory based-PR condition, the judgment task instructions were read after completing the summary table.

The cues that depicted the event states were different from those used in Experiments 1 and 2. The cues were still asymmetric (Allan, 1993) but they were displayed adjacent to a line that partitioned the right and left halves of the slide. The input and output variables depicted the relation between Variables X and Y. Variable X, if present, was depicted to the left of the partition, while Variable Y, if present, was depicted to its right. Microsoft PowerPoint software displayed the slides via a PC compatible LCD projector.

Measure and Items

Participant estimated the frequencies of causal information, judge the relative likelihood of events, indicate their contingency judgments on a line scale, and indicate

---

[8]The four coats of a traditional deck of cards were assigned to one of the four forms of causal information (i.e., ♦, ♣, ♠, and ♥ for *cells a*, *b*, *c*, and *d*, respectively). Cards appropriately representing each event state was entered into a hat and sampled without replacement. The order by the cards were removed from the hat determined the order of events in the forward condition.

the certainty in their estimates. Cell estimates, judgments of the relative likelihood, and the certainty ratings were recorded as explained in Experiment 1.

Contingency judgments were recorded on a line judgment scale similar to that used in Experiment 2. However, the labels and ticks were removed with the exception of the mid- and endpoints. The line judgment scale ranged from +100 (*much more likely*) to -100 (*much less likely*) and read as follows:

> On the scale below ranging from +100 to -100, draw a mark on the line that best characterizes the effect of Variable X on Variable Y.
>
> +100 0 -100
>
> |_______________________________|_______________________________|

Upon completing the line judgment scale, participants were asked to state the intended numeric value and sign of their line judgment.

Procedure

At the start of the experiment, participants were given a packet of materials. The experimenter began the session by reading the following instructions:

> In this experiment you will be presented with two sequences of slides containing abstract information. The first sequence will include 8 slides and will be used as a practice set to familiarize you with the experimental procedure. The second problem set will include 24 slides. The length of this experiment is expected to be a little over an hour. We ask that you do not attempt to work ahead of the rest of the group. Please remember to enter the last six-digits of your social security number in the top right hand corner of each sheet. This ID number will be used to keep information anonymous and confidential.
>
> Throughout this experiment, the slides will present information in one of the following four event forms.
>
> Variable X and Variable Y may both be present: [Insert Cell a]
>
> Variable X may be present while Variable Y may be absent: [Insert Cell b]
>
> Variable X may be absent while Variable Y may be present: [Insert Cell c]
>
> Variable X and Variable Y may both be absent: [Insert Cell d]

> These four forms of information can all be represented in and summarized in the following table: [Insert Summary Table]
>
> After viewing the slides, your task will be to fill-in the summary table with the frequency of each of the event forms shown during the sequence. Please free your hands of any pens or pencils during the presentation of the slides. The experimenter will inform you when it is okay to use your pen or pencil to fill in the summary table. The first will be a practice set consisting of 8 events, while the second problem sets will contain 24 events.
>
> If you have any questions concerning the procedure described thus far, please ask them now. This is not a competition and you should not talk to, consult with or observe the work of others. You will have 10 minutes to complete each summary table and the follow-up questions.

Each packet also included the judgment task instructions that informed participants of the later judgment task, which read as follows:

> Given that the two variables are presented together, a possible relationship may exist between the two variables. In the sequence of slides, it may be the case that Variable Y is either more likely, just as likely, or less likely to occur in the presence of Variable X. Following the completion of the summary table, your task will be to form a judgment of the strength and direction of the relation between Variable X and Variable Y.
>
> To asses your judgment of the relation between Variable X and Variable Y several follow-up questions will be used. Among other things, you will be asked to judge the relation between Variable X and Variable Y on a scale ranging from +3 to -3 (i.e., as shown below) and to mark a line on a scale corresponding to the strength and direction of the relation. In the completion of the follow-up questions you are free to view the contents of the summary table.
>
> Examples of the rating scales:
>
> Example A: [Insert judgment of relative likelihood scale]
>
> Example B: [Insert line judgment scale]
>
> If you have any questions concerning the procedure described thus far, please ask them now. This is not a competition and you should not talk to, consult with or observe the work of others. You will have up to 10 minutes to complete each summary table and the follow-up questions.

The judgment task instructions were inserted into the packet at points to accommodate the three experimental conditions. The consent form, demographic items, and closing of the experiment were conducted in the same manner described in Experiment 1.

Results

Cell Deviations

As explained in Experiments 1 and 2, the discrepancies between participants' estimates and the actual cell frequencies were used to create three variables that provide different information about the data: mean absolute cell deviations, mean signed cell deviations, and signed unit errors (A more detailed discussion of the differences between the three forms of data can be found in the Results section of Chapter I).

Mean Absolute Cell Deviations. To determine whether any differences existed in the magnitude of errors between cells, a series of paired-samples *t* tests were performed on the mean absolute cell deviations (top panel of Table A17). Previous research suggests that the increase in the magnitude of errors across cells is not a robust effect (Experiments 1 and 2; Shaklee & Mims, 1982). Therefore, two sets of test compared differences in the mean absolute cell deviations. The first set of tests assessed differences between *cell a* and each of the three remaining cells, while the second set assessed differences between the top and bottom rows of the summary table. Tests were conducted on the data when collapsed across all conditions and separately for each judgment task condition.

The first set of tests did not reveal any significant differences between cells. When data were collapsed across all conditions, the analysis revealed that the mean absolute cell deviation for *cell a* was not different than those obtained for either of the remaining three cells, *cell b*, $t(76) = 1.18$; *cell c*, $t(76) = -1.73$; *cell d*, $t(76) = -0.94$; all $ps \geq .09$. In the on-line condition, the mean absolute cell deviation for *cell a* was not different from those obtained for either of the remaining three cells, *cell b*, $t(25) = 0.68$;

cell c, $t(25) = -1.85$; *cell d*, $t(25) = -.75$; all *ps* $\geq .08$. In the on-line condition and when data were collapsed across all conditions, the difference in the mean absolute cell deviation between *cells a* and *c* approached but failed to reach conventional significance (i.e., $p = .09$ when data were collapsed; $p = .08$ in the on-line condition). In the memory based-PE condition, the mean absolute cell deviation for *cell a* was not different from those obtained for the remaining three cells, *cell b*, $t(27) = 1.10$; *cell c*, $t(27) = 0.00$; *cell d*, $t(27) = -0.56$; all *ps* $\geq .29$. In the memory based-PR condition, the mean absolute cell deviation for *cell a* was not different from those obtained for either of the remaining three cells, *cell b*, $t(22) = 0.00$; *cell c*, $t(22) = -1.31$; *cell d*, $t(22) = -0.37$; all *ps* $\geq .20$.

The second set of tests only indicated a significant difference for one of the analyses. When the data were collapsed across all conditions, the difference between the cells in the top row and bottom row was significant, $t(76) = -2.60$, $M_{SE} = 0.09$, $p = .01$. However, the difference between cells in the top and the bottom rows did approach statistical significance in the on-line condition, $t(25) = -1.94$, $M_{SE} = 0.169$, $p = .064$. All other *ps* were $\geq .14$.

<u>Mean Signed Cell Deviations</u>. To determine whether any differences existed in the magnitude of over- or underestimations, the mean signed cell deviations were examined (see bottom panel of Table A17). The current investigation has reported evidence that the mean signed cell deviations are positive for *cell a,* negative for *cells b* and *c*, and either negative (Experiment 1) or positive (Experiment 2) for *cell d.* Therefore, the paired-samples *t* test assessed differences between *cell a* and each of the three remaining cells. The planned comparisons were conducted on the data when collapsed across all conditions and separately for each judgment task condition.

For the most part, the difference in the mean signed cell deviations was significant when *cell a* was contrasted with *cells b* and *c*. When collapsed across conditions, the mean signed cell deviation for *cell a* was significantly different than those for *cells b* and *c*, *cell b*, $t(76) = -3.29$, $M_{SE} = 0.16$; *cell c*, $t(76) = -4.45$, $M_{SE} = 0.24$, all *p*s < .05. However, the mean absolute cell deviations for *cells a* and *d* were not significantly different from one another, $t(76) = 0.00$, $M_{SE} = 0.16$, $p = .99$. In the on-line condition, the mean signed cell deviation for *cell a* was again significantly different than those for *cells b* and *c*, $t(25) = -2.82$, $M_{SE} = 0.19$; $t(25) = -2.87$, $M_{SE} = 0.40$, all *p*s < .05. The mean absolute cell deviations for *cells a* and *d* were also not significantly different from one another, $t(25) = 0.00$, $M_{SE} = 0.22$, $p = .99$. However, the memory-based-PE condition failed to produce significant difference between *cells a* and those for *cells b* and *d*, *cell b*, $t(27) = -1.41$, $M_{SE} = 0.35$; $t(27) = 0.48$, $M_{SE} = 0.37$; all *p*s ≥ .17. The mean signed cell deviations for *cell a* in the memory-based-PE condition was significantly different than that obtained for *cell c*, $t(27) = -2.19$, $M_{SE} = 0.46$, $p < .05$. In the memory based-PR condition, the mean signed cell deviation for *cell a* was again significantly less than those for *cells b* and *c*, *cell b*, $t(22) = -2.65$, $M_{SE} = 0.26$; *cell c*, $t(22) = -2.74$, $M_{SE} = 0.40$, all *p*s < .05. The mean absolute cell deviations for *cells a* and *d* were again not significantly different from one another, *cell d*, $t(22) = -1.16$, $M_{SE} = 0.19$, $p = .26$.

Signed Unit Errors. Signed unit errors were calculated, as explained in Experiment 1. The percentage of participants that over-, under-, or correctly estimated the cell frequencies are shown in Tables A17 and A18. Table A18 displays the percentages of each type of signed unit error when collapsed across all experimental conditions (top panel), for the forward condition (middle panel), and for the backward condition (bottom panel). A t test indicated that there was no difference in the percentages of correct estimations in the event sequence conditions, $t(6) = -0.73$, $Diff_{SE} = 5.37$, $p = .46$.

Table A19 displays the percentages of each type of signed unit errors separately for the judgment task conditions: on-line judgment condition (top panel), memory based-PE condition (middle panel), and memory based-PR condition (bottom panel). Several t tests compared the percentages of correct estimations between judgment task conditions. There was no difference in the percentages of correct estimations in the on-line and memory based-PE conditions or between the two memory based conditions, on-line vs. post-encode, $t(6) = 0.62$, $Diff_{SE} = 5.12$; post-encode vs. post-recall, $t(6) = 0.50$, $Diff_{SE} = 3.88$, all $ps \geq .13$. However, the percentage of correct estimations in the on-line and memory based-PR condition were significantly different from one another, $t(6) = 1.39$, $Diff_{SE} = 3.68$, $p < .05$.

Mean signed unit errors were also calculated because they permit can serve as an index of participants' tendency to over- ($M > 0.00$) or underestimate ($M < 0.00$) cell frequencies. Means are also displayed in Tables A17 and A18. In all of the reported analyses, participants were more likely to underestimate confirming cases and overestimate disconfirming cases.

Judgments

Judgments of Relative Likelihood. Table A20 lists the frequency and cumulative percentage of participants that selected each option on the judgment of relative likelihood scale. The sign of participant's responses were used to determine judgment accuracy, as previously done for the rule-analysis technique (See Experiments 1 &2). As displayed in

Table A20, more than 50% of the participants correctly judged the problem set using the judgment of relative likelihood scale.

Line Judgments of Contingency. The value of line judgments were determined by dividing the distance of each response from the scale's midpoint by the total distance of the scale's corresponding endpoint (i.e., 0 to +100 positive responses or from 0 to -100 for negative responses). Negative line judgments (i.e., between -100 and 0) were multiplied by negative one to distinguish them from positive judgments. All positive line judgments were considered correct (i.e. 74.7% of the total responses), while those that were either negative or equal to zero were considered incorrect (i.e., 16 & 9%s of the total responses, respectively).

Numeric Estimate of Contingency. Judgments accuracy was also determined according to participants' numeric estimates of the strength and direction of the contingency. Again, all positive numeric estimates were considered correct (i.e. 71% of the total responses), while those that were either negative or equal to zero were considered incorrect (i.e., 18 & 11%s of the total responses, respectively).

Inter-Judgment Agreement. Pearsons correlation coefficients (r) were calculated on participant's responses on the three judgment scales. The resulting r values are displayed in Table A21 across all conditions and for each judgment task condition. The correlations between responses on each of the judgment scales were all significant at the $p < .01$ level. Due to the design of the current experiment, r values could not be calculated between judgments and Δp[9].

---

[9] The design of the current experiment only consisted of one problem set. For Pearsons correlation coefficients (r) to be calculated, both variables must vary to some extent. The inclusion of only one contingency (Δp) in the current experiment will not permit the calculation of r because the value of the contingency will be constant. This problem is typically overcome by requiring that participants judge

Judgment accuracies, as defined by the three judgment scales, were compared to determine whether of the judgment scale impacted judgment competency. Thirteen instances were identified in which a discrepancy existed between judgments on either of the three judgment scales (see Table A22).

Event Sequence

A multivariate analysis of variance (MANOVA) was used to determine whether event sequence had an effect on either of the dependent measures reported above. There were no reliable effects found for any of the reported variables (all *ps* $\geq$ .25).

Discussion

Cell Deviations

Mean Absolute Cell Deviations. Previous research has suggested that the differential pattern in the magnitude of errors is not a robust effect (Experiments 1 and 2; Shaklee & Mims, 1982). Therefore, the lack of significant differences between the mean absolute cell deviations is not of great concern. However, there was no evidence that the magnitude of errors increased across cells in any of the reported analyses (see Table A17). The current experiment only had one problem set, while previous experiments have all contained 12 problem sets (i.e., Experiments 1 and 2; Shaklee and Mims,1982). Therefore, it might have been that the differential pattern in the magnitude of errors emerges after repeated exposure to the experimental procedure (see Experiment 1 for a similar argument).

The failure to obtain the differential pattern in the magnitude of errors in either condition will not permit any conclusions about its likely source, as planned. If the

---

multiple contingencies or, more appropriate for this design, requiring that participants judge different contingencies.

differential pattern only occurs after repeated exposure, as suggested by the current data, the proposed procedure can not determine its source. It is not possible to contain cell estimates from multiple problem sets in which subjective cell importance only biases the encoding or recalling of cell frequencies.

Mean Signed Cell Deviations. The planned comparisons conducted on the mean signed cell deviations categorized the effect well. However, the signs of the first three cells of the summary table (i.e., *cells a*, *b*, and *c*) were opposite of what was predicted. The data indicated that the magnitude of underestimations for confirming cases were greater than that for overestimations and the exact opposite was true for disconfirming cases (see bottom panel Table A17). Again, the current experiment only included one problem set which explain the different trend obtained in the mean signed cell deviations.

Mean Signed Unit Errors. The means of the percentages of correct estimations, calculated across the four cells of each condition, were examined to determine whether differences existed between the experimental conditions. No differences in the percentages of correct estimations existed between the two event sequence conditions, which suggest that the order of presenting event states had no effect on the likelihood of correctly estimating cell frequencies.

The rank order of the mean percentage of correct estimations in the judgment task conditions was as follows: on-line (M = 69.22, SD = 7.03), memory-based-PE (M = 66.08, SD = 7.45), and memory-based-PR conditions (M = 69.22, SD = 7.03). It may be that knowledge of the judgment task prompted purposeful encoding and recalling, which resulted in an advantage when estimating cell frequencies. Participants' knowledge of the judgment task might have heightened the perceived importance of the recall task to ensure satisfactory performance on the later judgment task. Participants in the on-line judgment condition purposefully encoded and recalled the event states, while those in the memory-based-PE condition only purposefully recall them. Participants in the memory-based-PR did not purposefully encode or recall the event states, which might explain its ranking with respect to the judgment task conditions. If purposefully encoding and recalling do separately improve the accuracy of cell estimates, then it could account for the rank order obtained across conditions (see Saitoh, 1981 for comments on the benefits of meaningful learning).

The mean signed unit errors were also examined to assess participants' tendency to over- or underestimate cell frequencies. The mean signed unit errors indicated that participants were more likely to underestimate confirming cases and to overestimate disconfirming cases, which was also the opposite of what was predicted (see Mean Signed Cell Deviations in Results). The directions (i.e., signs) of the mean signed unit errors and mean signed cell deviations were identical in the current experiment. Again, the current experiment only included one problem set which might explain the different trend obtained in the mean signed unit errors.

Inter-Judgment Agreement. Bivariate correlations were performed to assess the extent that participants' judgments were similar on the judgment scales. The r values between line judgments and numeric estimates of contingency were stronger than all other correlations ($r \geq .97$). High levels between the line and numeric estimates might have occurred because the latter were judgments of the former. The relative likelihood scale was highly correlated with both the line and numeric estimates of contingency but to a lesser extent than they were correlated with one another ($.48 \geq r \geq .78$). The lower r values between the relative likelihood scale and the other scales suggest that participants did not use it as appropriately to judge contingency. Judgment accuracy, determined by the sign of participants' judgments, also suggested that the relatively likelihood scale was not as appropriately used as the line judgment and numeric estimate scale (i.e., 56, 71, and 75%s of participants' judgments were deemed correct, respectively). On several occasions responses on the judgment scales were inconsistent with one another (i.e., judgments made by the same participants contained conflicting signs; see Table A22). Differences in judgment accuracy between scales suggest that individuals do perceive them as asking different questions (see Shaklee & Tucker, 1980) and that the line judgment and numeric estimate scales are more appropriately used to judge contingency.

Conclusions

The current experiment set out to replicate the differential pattern in the magnitude of errors and determine whether it was due to biased encoding or recalling of causal information. The current experiment failed to produce a differential pattern in the magnitude of errors in any of the judgment task conditions. A re-analysis of Experiment 2's data, reported here for the first time, suggested that the differential pattern in the magnitude of errors is due to averaging data across multiple problem sets. As displayed in Table A23 (see note), the differential pattern was only present in 4 of the 12 problem sets. The emergence of the differential pattern only when the data are averaged is consistent with the notion that it is not a robust effect.

The current experiment failed to produce a differential pattern in the three judgment task conditions because it contained only one problem set. Therefore, the current design cannot determine whether subjective cell importance influences the differential pattern in the magnitude of errors. In the current design, cell estimates can only be obtained once because subjective cell importance will likely influence of additional estimates. However, other manipulations might provide insights into whether subjective cell importance influences the differential pattern. Experiments typically consists of measures that ask questions that direct participant's attention to *cell a* when judging contingency (e.g., Benassi, Knoth, & Mahler, 1985; Dennis & Ahn, 2001; Kao & Wasserman, 1993; Lopez et al., 1998; Mandel & Lehman, 1998; Shaklee & Mims, 1982; Ward & Jenkins, 1965). Several experiments have demonstrated that redirecting participant's attention to a different cell alters participants' beliefs about cell importance accordingly (Crocker, 1982; White, 2003). If subjective cell importance impacts the magnitude of errors, then shifting participants' attention should also alter the differential pattern.

Differences in the mean signed cell deviations across cells were also not as expected. The magnitude of underestimations was greater than the overestimations for confirming cases, while the exact opposite was true for the disconfirming cases. The signs for *cells a*, *b*, and *c* were opposite of what was found in Experiments 1 and 2, which might be accounted for by several factors. First, the trend might have been due to the use of only one problem set. A similar argument was also made for the mean absolute cell deviations. Second, it might be that the make-up of the contingency resulted in the following pattern. For example, the judged contingency was defined by the following frequencies of causal information: 10*a*, 4*b*, 2*c*, 8*d*. The majority of the frequencies were confirming cases, which were consistently underestimated in the current experiment. It may be that participants are more likely to underestimate cells that are frequent and overestimate cells that are relatively infrequent.

Experiment 2's data was re-analyzed to determine whether the actual cell frequencies influenced the tendency to over- or underestimate cell frequencies, reported here for the first time. The re-analysis indicated that the magnitudes of underestimations were greatest for high cell frequencies (i.e., if $f > 5$, then $M > 0.00$) and that the magnitudes of underestimations were greatest for those that are low (i.e., if $f > 5$, then $M > 0.00$). When means were calculated for each problem set, the sign of most cells could be predicted by assessing its frequency (i.e., 43 of 48 cases were predicted; see Table A24). The mean of the cell frequencies that defined the problem sets increased across *cells a*, *b*, *c*, and *d* in both Experiments 1 and 2 (5.08, 5.50, 6.25, & 7.17 in Experiment 1; 5.33, 5.42, 6.00, & 7.07 in Experiment 2, respectively). Therefore, the increase in the mean cell frequencies across *cells a*, *b*, *c*, and *d* can account for the differential pattern in the mean signed cell deviations. The same argument can be made for the differential pattern of errors in the mean signed unit errors because their signs have, for the most part, been the same across experiments.

To address the possibility that the increase in the magnitude of errors might also be accounted for by an increase in the magnitude of errors, the mean absolute cell deviations from Experiment 2 were also re-analyzed (Table A23). In only 5 of the 12 problem sets was the magnitude of errors lowest for the most infrequent cell and only on three occasions was it the second lowest cell that had the second lowest frequency (see Table A23 note). Therefore, it is unlikely that the cell frequencies also accounted for the increase in the mean absolute cell deviations. In addition, the mean cell frequencies increased across all four cells and the magnitude of errors for *cell d* was never found to be higher than that for *cell c* (cf. Shaklee & Mims, 1982).

Last, the current experiment also provided evidence that the structure of the judgment scale might influence the obtained judgment. There has been limited research dedicated to the impact of judgment scales on the likelihood of obtaining accurate judgments of contingency (Arkes & Rothbart, 1985; Shaklee & Tucker, 1980). If judgment scales do affect judgment accuracy, then further questions must also be asked. Researchers might wish to determine the judgment scale that is most likely to produce veridical judgments. Researchers might also wish to identify the judgment scale that is interpreted by participants as the equivalent to definition used in the contingency judgment literature.

In closing, the failure to support both the first and second proposed hypotheses was beneficial to the current investigation. As a result, it became evident that the differential pattern in the magnitude of errors is a by-product of averaging data. It also became evident that the differential pattern in the mean signed cell deviations and the mean signed unit errors are influenced by the cell frequencies used to define the problem sets. The third hypothesis, however, was supported and it suggested that different judgment items can be interpreted by participants as asking different questions that warrant vastly different responses. Although the majority of the current's experiments were null, they contributed greatly to the current investigation and has benefited this line of research.

CHAPTER IV

EXPERIMENT 4: THE SURPRISE ELEMENT

Research comparing on-line and memory-based judgment is limited in the number of problem sets that can be used to facilitate between-groups comparisons. Only one memory-based judgment can be obtained from each participant because additional inferences are likely to be on-line judgments. On-line judgments occur, by definition, when the individual is aware of the judgment task and updates his or her judgment as the relevant information is encountered. Memory-based judgments occur when the relevant information is encoded for another purpose; therefore, it must be recalled to form a judgment after the fact (Arkes & Rothbart, 1985; Hastie, Park, & Weber, 1984).

Experiment 3 compensated for the limited number of judgments with a practice problem set and practice slide. The practice problem set consisted of 8 event states (i.e., 3*a*, 2*b*, 2*c*, 1*d*) and the practice slide re-introduced them, in the same order, on a timeline. As the event states completed the once blank timeline, the cells of an adjacent summary table were simultaneously updated to reflect the causal information that had been introduced. At the end of the practice slide, the cells of the summary table were consistent with that which defined the practice problem set.

The practice problem set and practice slide were intended to familiarize participants with the recall task. Task familiarity was a concern because of the limited number of problem sets can be used[10]. In retrospect, the practice problem set and practice slide may have suggested a means by which to remember the cell frequencies. Therefore, the practice problem set and practice slide might also account for the absent differential patterns in Experiment 3 (see Experiment 3 for additional arguments).

A more elaborate technique will be used in the current experiment to ensure task familiarity that does not also suggest a means by which to complete the task. Participants in the memory-based and on-line conditions will view 5 problem sets. Participants in the memory-based condition will be told that the experiment is concerned with their ability to recall the cell frequencies, while those in the on-line condition will be told that it is concerned with their ability to judge the relation between events. After observing the fourth problem set, participants in both conditions will, for the first time, be informed of their second task. Participants in the memory-based condition will judge the relation between the events, while those in the on-line condition will be asked to recall the cell frequencies. This procedure is also limited in the number of problem sets that can be used to facilitate between-groups comparisons; however, participants should be familiar with the experimental task when the data are obtained.

Two hypotheses are proposed in the current experiment. First, it is expected that there will be a stronger relationship between judgments and inferred contingency ($\Delta p'$) than between judgments and actual contingency (Δp) in both judgment task conditions. Both Δp and $\Delta p'$ are calculated using the same equation; however, actual cell frequencies

[10] Two observations from previous experiments also suggested that some percentage of participants might not have been completely familiar with the experimental task upon starting the first problem set. First, on occasions participants have asked questions about the experimental task after having viewed the first problem set. Questions were asked even though the experimenter previously had asked, prior to starting the first problem set, whether anyone had any questions. Second, several participants in previous experiments did not provide estimates for the first problem set (i.e., 1 and 4 participants in Experiments 1 and 2, respectively). Although only a speculation, it might be that participants that did not provide estimates for the problem set were not familiar with the task after viewing it.

serve as inputs for Δp and participants' estimates of the cell frequencies are used for $\Delta p'$. The Δp statistic is frequently used as an index of the extent that variables are related, while I have argued that $\Delta p'$ is an index of the perceived contingency (Note that $\Delta p'$ oftentimes differs from Δp due to memory error; see Experiments 1-3). Second, it is expected that there will be a stronger relationship between participants' judgments and $\Delta p'$ in the memory-based condition than in the on-line condition.

Method

Participants

Ninety-six students (20 males and 75 females, mean age = 19.43) enrolled in lower-level psychology courses at the University of New Hampshire participated in the current experiment. All participants received course credit for their participation. Sixty and 24%s of the participants were freshman and sophomores, respectively. Sixteen percent were upperclassmen (i.e., 12.6 and 3.4%s were juniors and seniors, respectively).

Design

The design consisted of a within-subjects factor (Cell) and a between-subjects factor (Type of Judgment). Cell (4 levels) refers to participants' estimates of the actual cell frequencies, which are displayed in Table A25 with the Δp for each problem set. Type of Judgment (2 levels) refers to the judgment task that distinguished the two experimental conditions. Abstract and contextual cues represented the event states, which were identical to those used in Experiment 2. Microsoft PowerPoint software displayed the slides via a PC compatible LCD projector at a rate of 1 event state every 4 s.

Measure and Items

Participants, when appropriate, estimated the frequencies of causal information (i.e., the recall task), judged the contingency (i.e., the judgment task), and indicated their certainty in the previously completed recall task and/or judgment task. All measures were identical to those used in Experiment 2.

Procedure

Participants received a packet that contained the instructions, consent form, five multiple-item measures, and debriefing form. The consent form was administered after the instructions; the instructions in the memory-based condition read as follows:

> In the current experiment you will view five Problem Sets that portray the Event States of two variables. Event States provide distinct information about the absence or presence of the two variables. Consider the case in which two variables correspond to Cows (Variable 1) and Grass (Variable 2). It may be the case that either the Cows and Grass are present on a farm (Event State 1 – Cell 1) or that Cows are present and there is No Grass (Event State 2 – Cell 2), It might also be the case that there are No Cows and Grass is present (Event State 3 – Cell 3) or that there are No Cows and No Grass (Event State 4 – Cell 4) on the farm. The five Problem Sets will each consist of 24 Event States and will be defined by any combination of the four Event States, as described above.
>
> [Insert Summary Table]
>
> During the viewing of each Problem Set, you must put down your writing utensil and rest your hands. After viewing each Problem Set, you will be prompted to recall the number of instances that each of the four Event States occurred. With your pen or pencil, please enter the number of times that each Event State occurred in the appropriate cells of a blank table, like the one above.
>
> If you have any questions concerning the procedure described thus far, please ask them now. This is not a competition and you should not talk to, consult with, or observe the work of others.

The instructions for the on-line condition differed from those listed above. The differences informed participants of the judgment task instead of the recall task and read as follows:

> During the viewing of each Problem Set, you must put down your writing utensil and rest your hands. After viewing each Problem Set, you will be prompted to

judge the relationship between the set of Event States. With your pen or pencil, please mark on the provided the extent that the two variables are related as portrayed by the event states.

[Insert Contingency Scale]

In the two experimental conditions, the task order for Problem Sets 4 and 5 differed. Participants in the on-line condition viewed the problem set, completed the recall task, judged the extent that the two variables are related, and then completed the two certainty ratings. Participants in the memory-based condition viewed the problem set, completed the judgment task, estimated the frequencies of causal information, and then completed the two certainty ratings.

The demographic items and closing of the experiment were conducted in the manner described in Experiment 1.

Results

Judgments

Bivariate Correlations. Pearsons correlation coefficients (r) were calculated between judgments of contingency, Δp, and $\Delta p'$ when the data were collapsed across all conditions and separately for each judgment task condition (see Table A29). When the data were collapsed across all conditions, judgments were significantly correlated with Δp, $r = 0.37$, $n = 327$, $p < .01$. Judgments of contingency were also significantly correlated with $\Delta p'$, $r = 0.29$, $n = 153$, $p < .01$. The difference between the two previously mentioned correlations was not significant, $t(327) = 1.88$, $p = .97$. Last, Δp and p' were significantly correlated with one another, $r = 0.66$, $n = 330$, $p < .01$.

A similar pattern was found in the on-line condition. Judgments were again significantly correlated with Δp, $r = 0.25$, $n = 232$, $p < .01$. Judgments of contingency were also significantly correlated with $\Delta p'$, $r = 0.62$, $n = 90$, $p < .01$. The difference between the two previously mentioned correlations was also significant, $t(229) = -10.45$, $p < .05$. Last, Δp and p' were also significantly correlated with one another, $r = 0.73$, $n = 95$, $p < .01$.

A similar pattern was also found in the memory-based condition. Judgments were again significantly correlated with Δp and $\Delta p'$, Δp, $r = 0.67$, $n = 95$; $\Delta p'$, $r = 0.55$, $n = 93$; all *p*s $< .01$. The difference between the two previously mentioned correlations was not significant, $t(92) = 2.11$, $p = .98$. Last, Δp and p' were also significantly correlated with one another, $r = 0.73$, $n = 95$, $p < .01$.

To test the second hypothesis, the difference in the correlations between judgments and $\Delta p'$ in the memory-based and on-line conditions was tested for significance. The difference between judgments and $\Delta p'$ in the judgment task conditions was not significant, $z_{diff} = .71$, $p = .76$.

Judgment Accuracy. Mean judgments are displayed for all problem sets in Table A25. Only the resulting mean judgments for Problem Sets 4 and 5 are discussed because they were the only two judged by all participants. Mean judgments for Problem Set 4 in the memory-based condition ($M = .27$, $SD = .26$, $Mdn = .30$, Mode = .30) and on-line condition ($M = .21$, $SD = .36$, $Mdn = .20$, Mode = .00) were both appropriately positive. Mean judgments for Problem Set 5 in the memory-based condition ($M = -.28$, $SD = .34$, $Mdn = -.30$, Mode = .00) and on-line condition ($M = .27$, $SD = .36$, $Mdn = -.30$, Mode = -.30) were both appropriately negative.

Judgment accuracy was also determined by the sign of participants' resulting judgments, as done for the rule-analysis technique (see Shaklee, 1983). Problem Set 4 was correctly judged by 60% of the participants in the memory-based and 75% and those in the on-line judgment conditions. Problem Set 5 was correctly judged by 64% of the participants in the memory-based and 77% of those in the on-line judgment conditions.

Cell Deviations

The discrepancies between participants' estimates and the actual frequencies of causal information were used to create three variables that provide different information about the data: mean absolute cell deviations, mean signed cell deviations, and signed unit errors (see Chapter I for more details).

Mean Absolute and Signed Cell Deviations. There were no hypotheses or planned comparisons proposed for the mean absolute or signed cell deviations (see Tables A26 and A27, respectively). Therefore, the data will be reviewed in the Discussion.

Mean Signed Unit Errors. Table A28 displays the percentage of correct estimations for each problem set, separately for each cell. Figure B2 displays the trend in the mean percentage of correct estimations (listed as % Correct), calculated across the four cells, for each problem set. The trend across problem sets was not expected; therefore, paired-samples *t* tests were conducted to assess the differences across problem sets post hoc (i.e., Problem Set $1 \neq 2$, $2 \neq 3$, $3 \neq 4$, $4 \neq 5$). In the memory-based condition, differences in the mean percentage of correct estimations between Problem Sets 1 and 2, 3 and 4, and 4 and 5 were significant, 1 vs. 2, $t(3) = -3.51$, $M_{SE} = 4.17$; 3 vs. 4, $t(3) = 6.70$, $M_{SE} = 4.53$; 4 vs. 5, $t(3) = -4.53$, $M_{SE} = 3.70$; all $ps < .05$. No other differences in either condition reached statistical significance.

Table A28 also displays the mean signed unit errors from the current experiment, which will also be reviewed in the Discussion.

Discussion

Judgments

Bivariate Correlations. As was the case in Experiment 2, the correlations between judgments and Δp as well as between judgments and $\Delta p'$ were all significant. To test the first hypothesis, the difference in the correlations between judgments and Δp and judgments and $\Delta p'$ was calculated. The difference was only significant the on-line condition, in which the relation between judgments and $\Delta p'$ (r = .62) was significantly stronger than that for judgments and Δp (r = .25). Not only was there not a significant difference between the correlations in the memory-based condition, but the values were not in the expected direction. That is, the correlation between judgments and Δp (r = .67, $p < .05$) was slightly lower than that between judgments and $\Delta p'$ (r = .55, $p < .05$).

A number of factors may have contributed to the reported results. First, the task order in the memory-based condition differed from that used in previous experiments. Forming a judgment prior to declaring the estimates of causal information may have impaired cell accuracy (i.e., due to forgetting and/or interference). Evidence suggests that cell accuracy for Problem Set 4 in the memory-based condition was affected as a result of forming their judgment first. As displayed in Figure B2, the mean percentage of correct estimations increased across the first three problem sets; however, it decreased on Problem Set 4 (M = 19.48, SD = 3.10) to levels comparable to that on the first problem set (M = 19.70, SD = 1.93). It may be that the mental representation stored in memory was rapidly deteriorating as time passed. The deterioration might also have been exacerbated by the introduction of the judgment task [11]. Introducing the judgment task

[11] In the current and previous experiments, participants have been observed declaring their frequency estimates very quickly at the end of the problem set and being told to complete the multi-item measure.

prolonged the declaring of the frequency estimates, which may have resulted in the declared estimates being quite different from that which was used to form their judgment.

Judgment Accuracy. Judgment accuracy was assessed using two techniques. Mean judgments were calculated and the percentages of correct judgments were compared. Mean judgments in the on-line ($M = .27$, $SD = .26$) and memory-based conditions ($M = .21$, $SD = .36$) for Problem Set 4 were similar in sign and strength, which suggests that regardless of the judgment task condition participants arrived at similar judgments.

The data indicated that a slightly higher percentage of participants in the on-line condition correctly judged Problem Sets 4 and 5 (i.e., 75 and 77%s, respectively) than in the memory-based condition (60 and 64%s, respectively). I have argued that perhaps $\Delta p'$ is a more appropriate statistic from which to determine judgment accuracy, which seems even more appropriate for the memory-based condition. A re-analysis of the data obtained for Problem Set 4 indicated that five participants perceived (i.e., according to their $\Delta p'$) a negative contingency and provided a negative judgment. In addition, three participants perceived a noncontingent relation (i.e., $-.05 \geq \Delta p' \geq .05$) and judged it as noncontingent. The same analysis for Problem Set 5 indicated that three participants perceived a noncontingent relation (i.e., $-.05 \geq \Delta p' \geq .05$) and judged it as noncontingent. No participant perceived the relation as negative and judged it accordingly. Judgment accuracy improves in both the memory-based (i.e., 71 and 77% for Problem Sets 4 and 5, respectively) and on-line conditions (i.e., 81 and 67%s for Problem Sets 4 and 5,

---

Participant's eagerness to complete the measure may have been due to their own knowledge that the accuracy of their recall is greatly reduced as time passes.

respectively) if judgments consistent with $\Delta p'$ are also considered correct, which further supports the notion that it can be used to assess judgment accuracy.

Cell Deviations

Mean Absolute Cell Deviations. The lack of a general trend in the magnitude of errors might have been due to several factors (Table A26). First, it might have been that the current experiment did not contain enough problem sets for the differential pattern to emerge (see Experiment 3 for an extended argument). Second, the mean percentage of correct estimations was impaired by the introduction of the judgment or recall task during Problem Set 4. Introducing a novel task may have interfered with the emergence of the differential pattern. Last, judgments were not required for Problem Sets 1 through 3 in the memory-based condition. I have argued throughout this investigation that the differential pattern in the magnitude of errors might be due to beliefs about each cell's importance for judging contingency. Therefore, Problem Sets 4 and 5 in the memory-based condition are the only problem sets in which beliefs about subjective cell importance could have biased participants' cell estimates.

In Experiment 3, the increase in the magnitude of errors across cells was affected by an increase in the mean cell frequencies. The mean cell frequencies did not increase across cells in the current experiment: 6.2, 5.4, 7.8, and 5.2 for *cells a*, *b*, *c*, and *d*. Therefore, differences in the mean absolute cell deviations across cells might prove relevant. When data were calculated across the entire experiment, there was no evidence that the magnitude of errors increased as the mean cell frequencies increased (see last row of Table A26). The magnitude of errors did increase as cell frequencies increased within problem sets (i.e., in 20 of the possible 48 cases; see note in Table A26), which is extremely damaging to the results first reported by Shaklee and Mims' (1982). Future research is needed to determine whether the increase in the magnitude of errors is due to an increase in the mean cell frequencies[12].

Mean Signed Cell Deviations. There also were no proposed hypotheses for the mean absolute cell deviations (Table A27). As was the case with the data obtained in Experiments 2 and 3, the directions (i.e., signs) of the mean signed cell deviations were well predicted by the actual cell frequencies (i.e., 33 of 36 cells were correctly predicted; see note in Table A27). Participants underestimated cell frequencies to a greater extent when they were high (i.e., if $f > 5$, then $M > 0.00$) and overestimated them to a greater extent when they were low (i.e., if $f > 5$, then $M > 0.00$; see Table A27); this tendency will hereafter be referred to as the *cell estimation rule*. Ironically, the cell estimation rule was not successful predicting the sign of the mean signed cell deviations in the memory-based condition; that is, when calculated across problem sets (*M*s = 6.2, 5.4, 7.8, and 5.2 for *cells a*, *b*, *c*, and *d*). The cell estimation rule did predict the signs of the mean signed cell deviations when calculated across all problem sets in the on-line condition[13].

Signed Unit Errors. When the signed unit errors were calculated, three trends were identified in which the percentage of correct estimations increased across consecutive Problem Sets (Figure B2). There was an increase across the first three problem sets in the memory-based condition and, after a significant decrease between

---

[12] The problem sets in the current experiment were not designed to assess whether an increase in the absolute value of the cell frequencies resulted in an increase in the magnitude of errors. As a result, the difference in the mean cell frequencies between *cells b* and *d* was very small. A larger discrepancy in the mean cell frequencies between *cells b* and *d* would have permitted this hypothesis to be fully addressed.
[13] Note that the means averaged across all problem sets in the on-line condition were not based on five problem sets because estimates were only required following Problem Sets 4 and 5. Therefore, the mean cell frequencies in the on-line condition were different for *cells a*, *b*, *c*, and *d* (i.e., 6.5, 6.5, 5, and 6, respectively). This signs of the mean signed cell deviations were well predicted in the on-line condition by the cell estimation rule.

Problem Sets 3 (*M* = 50.03, *SD* = 8.82) and 4 (*M* = 19.70, *SD* = 1.93), it again increased in both the memory-based and on-line conditions. The increase in the percentage of correct estimations suggests that correctly estimating cell frequencies improves with practice, which is consistent with the notion that strategies are used to aid recall (see Experiments 1 and 2). The decrease in the percentage of correct estimations in the memory-based condition was argued above to have been due to either the introduction of the judgment task or the order of the two experimental tasks (see Bivariate Correlations in Discussion).

The mean signed unit errors were also calculated (Table A28). As with previous experiments, there was a high level of consistency between the signs of mean signed unit errors and the mean signed cell deviations. Therefore, the cell estimation rule could also predict the sign of the mean signed unit error. There was one discrepancy in the signs between the mean signed unit errors and mean signed cell deviations occurred for *cell d* in Problem Set 1, in which the cell estimation rule predicted a negative sign and the mean signed unit error was positive (*M* = .02, *SD* = .90). The high level of consistency between the two variables implies that when participants are more likely to overestimate cell frequencies their magnitudes are greater than the underestimations. The exact opposite can be said for the underestimations.

Conclusions

To summarize, contingency judgments were more highly correlated with $\Delta p'$ than with Δp. The one exception occurred in the memory-based condition, in which the correlation between judgments and Δp was not different from that between judgments and $\Delta p'$. As a result, the first hypothesis was partially supported.

The second hypothesis was clearly not supported. The difference in the correlations between judgments and $\Delta p'$ in the memory-based and on-line conditions was not significant. In fact, the difference was in the opposite direction from that which was expected. The results of the current experiment suggest that judgments are more highly correlated with memory in the on-line condition than in the memory-based condition, which is the exact opposite of what was predicted. However, the order in which the recall and judgment task were completed differed in the on-line and memory-based conditions. Perhaps if they were equated, the second hypothesis would have been supported.

The current experiment provided further support for the use of $\Delta p'$ to assess judgment accuracy. The percentages of judgment accuracy were also improved when the sign of $\Delta p'$ was used instead of Δp (cf. Shaklee & Mims, 1982; Shaklee & Tucker, 1980). It seems reasonable to assess judgment accuracy with what the individual remembers of the contingency ($\Delta p'$) rather than with the actual contingency (Δp). The use of $\Delta p'$ seems even more justified when time has elapsed between the viewing of the events and the forming of the judgment or when the frequencies of causal information are unknown (i.e., to the experimenter and to the individual). Future research should attempt to quantify the benefits of using $\Delta p'$ over Δp when assessing judgment accuracy (see Experiment 3 for a more detailed discussion).

The current experiment did not intend to address the impact of the actual cell frequencies on cell deviations and, as a result, strong conclusions can not be made. However, the mean absolute and mean signed cell deviations were both influenced by the actual cell frequencies. The influence of the actual cell frequencies on the cell deviations is independent of beliefs about subjective cell importance. To the best of my knowledge, no body of psychological research has documented tendencies to estimate frequencies in the manners reported in the current experiment. Therefore, the impact of each cell's frequency on the cell estimates may be an artifact of the experimental task.

Summary of Phases I and II: General Discussion

The current investigation has documented several important contributions to the contingency judgment literature. The following discussion summarizes the main contributions of the second phase of the current investigation, while mentioning related areas that will benefit from further research.

Cell Deviations

Mean Absolute Cell Deviations. The current investigation set out to determine whether participants differentially make errors when recalling cell frequencies, as suggested by Shaklee and Mims (1982). Initially, it was argued that the differential pattern in the magnitude of errors were due to beliefs about subjective cell importance. For example, the rank order of participants' beliefs about cell importance (i.e., $a > b > c >$ d) might have influenced the magnitude of errors when recalling cell frequencies (i.e., $a <$ $b < c <$ d). The current investigation never found that the magnitude of errors increased across all four cells (cf. Shaklee & Mims, 1982). However, the magnitude of errors did increase across the first three cells (see Experiments 1 & 2). Evidence suggested that the differential pattern in the magnitude of errors is a weak effect that is not promoted by beliefs about subjective cell importance, but by the cell frequencies used to define problem sets. In Shaklee and Mims' (1982) experiment and Phase I of the current investigation, the increase in the magnitude of errors was confounded by an increase in the mean cell frequencies. However, the mean cell frequencies did not increase across cells in Experiment 4 and, within problem sets, increases in the magnitude of errors tracked the increase in cell frequencies (see Table A26 note). Further research is needed. However, the evidence gathered thus far challenges the results first reported by Shaklee and Mims (1982).

Mean Signed Cell Deviations. Originally, there were no proposed hypotheses about the mean signed cell deviations and it was unknown whether they would provide any useful insights about the data. Data throughout the current investigation suggested that the actual cell frequencies influenced whether the magnitude of errors was greater for over- or underestimations. If the cell frequency was less than or greater than 5, then the magnitude of errors was greater for over- or underestimations, respectively. The influence of the actual cell frequency on the magnitude and direction of estimation errors was referred to as the cell estimation rule. The one limitation of the cell estimation rule is that it can not predict the sign of the mean signed cell deviations for cells that contain 5 events. If a cell contained 5 events the magnitude of errors were equally as likely to be either under- or overestimations.

The cell estimation rule might be explained by a ceiling and floor effect. The greater magnitude of overestimations for cell frequencies less than 5 might be explained by a floor effect. The range of possible estimates on the lower end is restricted because a negative frequency is not possible, which might permit the unrestricted range of the upper end (i.e., overestimations) to be of a greater magnitude and result in a positive mean signed cell deviation. While a real ceiling does not exist on the upper end for cell frequencies greater than 5, it could be argued that participants are aware that the problem sets are not exclusively defined by one event state. Therefore, knowledge of the distribution of events might similarly serve as a ceiling and restrict the upper-end of participants' estimates.

Mean Signed Unit Errors. The signed unit errors served two purposes throughout the current investigation. First, signed unit errors permitted an assessment of the percentages of estimates that were over-, under-, and correct estimations. The signed unit errors permitted an indication of the likelihood of errors when estimating cell frequencies, which was absent from the work reported by Shaklee and Mims (1982). Close to 50% of the total estimations were correct. Shortening the event state duration (Experiment 1) and introducing a surprise task (Experiment 4) decreased the percentage of correct estimations. Second, the mean signed unit error served as an index of the tendency to over- or underestimate cell frequencies. The signs of the mean signed unit errors and mean signed cell deviations were, for the most part, consistent with one another; however, occasionally their signs were different and on each of those occasions the latter variable was better predicted by the cell estimation rule. Therefore, the mean signed cell deviations are better characterized by the cell estimation rule. The tendency to over- or underestimate actual cell frequencies was systematic, which differed from that concluded by Smedslund (1963). The cell estimation rule, when applied to the data reported by Smedslund (1963; see Table II in Smedslund on pp. 168), could not predict the tendency to over- or underestimate cells in his experiment. However, numerous differences exist between the two procedures which might not warrant such a comparison.

Final Reflections on Cell Estimates. The current investigation demonstrates that cell estimates are influenced by a variety of situational factors. First, knowledge of the total number of event states influenced the sums of participant's estimates (see Experiment 2). Second, knowledge of the distribution of causal information is thought to influence the upper end of participant's estimates; that is, when cells contain more than 5 events (see Experiment 4). The two previously mentioned influences suggest that knowledge of the situation promotes cognitive reasoning that affects reports of that stored in memory. Similar types of reasoning might always aid memory and pure indications of that stored in memory are never obtained, especially when reporting on complex or ambiguous events. Similar plausible forms of reasoning have been found to influence discourse processing (e.g., Lea, 1995; Lea, Mulligan, & Walton, 2005), memory of pictures (e.g., Allport, 1954; Seamon, et al., 2002), and eyewitness testimonies (e.g., Tuckey & Brewer, 2003). Perhaps there is a natural tendency to incorporate logical reasoning into the interpreting of information stored in memory (see Bartlett, S. F. C., 2003; Von Hecker, 2004).

Memory and Judgment

The primary purpose of the current investigation was to illuminate the role of memory for judging contingency, which has not received much attention in the experimental literature. The current investigation documented evidence that warrants the continued investigation of the role of memory for judging contingency, which will be discussed below.

First, the current investigation suggests that contingency judgments are based on individuals' memory of the contingency. Contingency judgments were more highly correlated with $\Delta p'$, which uses estimates of cell frequencies, rather than Δp, which uses the actual cell frequencies. I argue that judgments are more highly correlated with $\Delta p'$

because participants crudely attempt to judge contingency from the frequencies of causal information. Participants' estimates of the cell frequencies often differ from that which actually occurred. Experiments 2 and 4 suggest that compensating for each participant's discrepant mental representation (i.e., calculating $\Delta p'$) improved the ability to predict contingency judgments; that is, with respect to the ability to predict judgments from Δp. In other words, errors in judgment can partially be accounted for by errors in memory. Therefore, it is reasonable to conclude that the information stored in memory is used to judge contingency.

Second, the current investigation suggests that judgments are formed by rule-based strategies or some rough approximation of them because the only difference between Δp, a rule-based strategy, and $\Delta p'$ are its inputs. Therefore, the higher correlation obtained when correlating judgments with $\Delta p'$, instead of Δp, suggests that judgments are formed in a manner that approximates the delta rule. Shaklee (1983) pointed out that the outputs of all of the rule-based strategies are highly correlated with one another. Therefore, suggesting that the current experiment provides evidence for one specific rule is not warranted. The marked improvement in the correlations when using $\Delta p'$ also suggests that some computational process is at work when judging contingency. Both the Rule-Based and Belief-Revision Models have computational components in their accounts of contingency judgment; therefore, the results are more in favor with their accounts than with the Associative Model.

Third, the current investigation suggests an alternative means by which to assess the accuracy of contingency judgments. There has been some debate as to the appropriateness of using a normative statistic to assess the accuracy of participants' judgments (Mandel & Lehman, 1998; Shaklee, 1983). I have argued that perhaps $\Delta p'$ is a more appropriate means by which to compare judgment accuracy. Further research is needed to quantify the benefits of using $\Delta p'$ over Δp; however, the arguments proposed here have been promising. For example, it has been argued that $\Delta p'$ might be more beneficial when the relevant information is not controlled or known to the experimenter. It may also be that causal knowledge, preconceived notions, stereotypes, illusory correlations, and biases are the result of inaccurate $\Delta p'$.

Final Reflections on Judgment and Memory. The current investigation has demonstrated that the role of memory for judging contingency deserves experimental attention. The issues discussed within the current investigation have investigated two possible applications of the role of memory for judging contingency. That is, the memory of causal information and the dependence of judgments on memory under different task conditions. There are many other components of memory that are relevant to contingency judgments. For instance, the frequency estimates were more than likely stored in short-term memory. While forming the contingency judgment is likely to occur in working memory, which many believe is different from short-term memory (Baddeley, 1995; Baddeley, 2000). The Belief-Revision Model holds that past causal knowledge impacts contingency judgments, which is more than likely stored in and retrieved from long-term memory. Therefore, all three major structures of memory (i.e., short-term, working, long-term memory) are worthy of investigation.

Future directions in the current line of research will explore the naturalistic applications of the current investigation's finding. For example, the contingencies used throughout the current investigation were defined by, on average, 24 event states.

Judgments made in naturalistic settings are often based on much larger samples. It must be determined whether the relationship between judgment and memory exist when frequencies are unknown or are high in quantity. Consider the relationship between memory and judgments in free-operant procedures. Free-operant procedures permit participants to respond on an operandum as many times as they wish throughout a set period of time and their responses either will or will not produce an outcome (Allan, 1993). High rates of responding are frequently observed using free-operant procedures (Benassi, Knoth, and Mahler, 1985; Wasserman, Chatlosh, & Neunaber, 1983). Therefore, it is unlikely that a participant would form a judgment on the actual frequencies of causal information. However, participants would be capable of estimating the percentage of each of the four forms of causal information. The delta rule produces the exact same output when the percentages of each event state are used as inputs instead of its actual frequencies[14]. It may be that participants use a similar process when judging contingency when the total number of frequencies are unknown or are abundant.

Future research should also address the major limitation encountered in the current investigation. In Experiment 4, the data did not indicate that judgments in the memory-based condition were more highly correlated with $\Delta p'$ than with Δp. The failure was argued to have been the result of switching the task order, which intended to rule out an alternative explanation for the results obtained in Experiments 1, 2, and 3. Switching the task order intended to demonstrate that declaring cell estimates before forming a judgment did not increase the likelihood that judgments were based on their frequencies,

---

[14] Consider the following contingency: 11*a*, 4*b*, 8*c*, 1*d*. The percentages of the causal information are .46, .33, .16, and .04, respectively. That is, if the frequencies are divided by the number of event that defined the problem set (i.e., 24). The output of the delta rule based on the percentages (.46/[.46+.33])- (.14/[.16+.04]) = -.22) is identical to its output when the frequencies are used (11/[11+4])-(8/[8+1]) = -.22).

which might not have occurred otherwise. Fortunately, there are other ways to reduce the likelihood that declaring cell estimates influences the following judgment. First, the order in which the event states are requested might be counterbalanced. The order in which cell estimates were declared might have suggested a ranking of their importance or a means to integrate them to judge contingency (see Wasserman, Dorner, & Kao, 1990 for a similar argument). Second, the request for estimates of the frequencies of causal information might be listed in sentence form (see Crocker, 1982). As argued earlier (see Introduction), the presence of the contingency table might elicit a greater tendency for participants to a use rule-based strategy when judging contingency. Third, upon declaring each frequency estimate they can immediately be removed from the participants' presence. Any combination of the above stated manipulations might compensate for the concern that motivated changing the task order in Experiment 4. The most logical progression of this line of research is to address the previously mentioned concerns, which will serve to strengthen the arguments stated throughout this investigation.

REFERENCES

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

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and nature of the response alternatives. *Canadian Journal of Psychology, 34*, 1-11.

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

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

Alloy, L. B., Crocker, J., & Tabachnik, N. (1980, August). *Depression and covariation judgments: Expectation based distortions in information search and recall.* Paper presented at the meeting of the American Psychological Association, Montreal, Canada.

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, N. H., & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Behavior, 2*, 379-391.

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

Baddeley, A. D. (1995). The psychology of memory. In A. D. Baddeley, & B. A. Wilson (Eds.), *Handbook of memory disorders* (pp. 3-25). Oxford, England: John Wiley & Sons.

Baddeley, A. D. (2000). Short-term and working memory. In E. Tulving, & F. I. M. Craik (Eds.), *Oxford handbook of memory* (pp. 77-92). London: Oxford University Press.

Baker, A. G., & MacKintosh, N. (1977). Excitatory and inhibitory conditioning following uncorrelated presentations of CS and UCS. *Animal Learning and Behavior, 5*, 315-319.

Baker, A. G., & MacKintosh, N. (1979). Preexposure to CS alone, US alone or CS and US uncorrelated: Latent inhibition, blocking by context or learned irrelevance? *Learning and Motivation, 13*, 391-416.

Baker, A. G., Murphy, R. A., & Vallée-Tourangeau, F. (1996). Associative and normative models of causal induction: Reacting to versus understanding cause. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *Causal learning* (pp. 1-45). San Diego, CA: Academic Press.

Baron, A., & Perone, M. (1998). Experimental design and analysis in the laboratory study of human operant behavior. In K. A. Lattal, M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 45-91). New York: Plenum Press.

Bartlett, S. F. C. (2003). Remembering: A study in experimental and social psychology. In M. P. Munger (Eds.), *History of psychology: Fundamental questions* (pp. 430-446). London: Oxford University Press.

Bennasi, V. A., Knoth, R. L., & Mahler, H. I. M. (1985). Detection of noncontingency in a free-operant situation. *Personality and Social Psychology Bulletin, 11*, 231-245.

Catena, A., Maldonado, A. Megias, J. L., & Bettina, F. (2002). Judgment frequency, belief revision, and serial processing of causal information. *The Quarterly Journal of Experimental Psychology, 55B*, 267-281.

Catena, A., Maldonado, A., & Candido, A. (1998). The effect of frequency of judgment and the type of trials on covariation learning. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 481-495.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior, 6*, 151-155.

Crocker, J. (1981). Judgment of covariation by social perceivers. Psychological Bulletin, 90, 272-279.

Crocker, J. (1982). Biased questions in judgment of covariation studies. *Personality and Social Psychology Bulletin, 8*, 214-220.

Crocker, J., & Taylor, S. E. (1978, August). *Theory-driven processing in the use of complex evidence*. Paper presented at the meeting of the American Psychological Association, Toronto, Canada.

Dennis, M. J., & Ahn, W. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory and*

Cognition, 29, 152-164.

Hastie, R. (1984). Causes and effects of causal attributions. *Journal of Personality and Social Psychology, 46*, 44-56.

Hastie, R. & Kumar, A. P. (1979). Person memory: Personality traits as organizing principles in memory for behavior. *Journal of Personality and Social Psychology, 37*, 25-38.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*, 258-268.

Hastie, R., Park, B., & Weber, R. (1984). Social Memory. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 151-212). Hillsdale, NJ; Erlbaum.

Hastie, R., & Pennington, N. (1989). Notes on the distinction between memory-based versus on-line judgments. In John N. Bassili (Ed.), *On-line cognition in person perception* (pp.1-17). Hillsdale, NJ: Erlbaum.

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

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

Kao, S., & Wasserman, E. A. (1993). Assessment of an Information Integration Account of Contingency Judgment of Subjective Cell Importance and Method of Information Presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1363-1386.

Lea, R. B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1469-1482.

Lea, R. B., Mulligan, E. J., & Walton, J. L. (2005). Accessing distant premise information: How memory feeds reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 387-395.

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

Maldonado, A., Catena, A., Candido, A., Garcia, I. (1999). The belief revision model: Asymmetrical effects of noncontingency on human covariation learning. *Animal Learning and Behavior, 27*, 168-180.

Maldonado, A., Martos, R., & Ramirez, E. (1991). Human contingency judgments in terms of previous controllability and actual contingency. *Quarterly Journal of Experimental Psychology, 43B*, 347-360.

Mandel, D. R., & Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General, 127*, 269-285.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Peterson, C. (1980). Recognition of noncontingency. *Journal of Personality and Social Psychology, 38*, 727-734.

Saitoh, S. (1981). Cue utilization in incidental learning of normal and retarded children. *Tohoku Psychological Folia, 40*, 1-5.

Seamon, J. G., Schlegel, S. E., Hiester, P. M., Landau, S. M., & Blumenthal, B. F. (2002). Misremembering pictured objects: people of all ages demonstrate the boundary extension illusion. *American Journal of Psychology, 115*, 151-167.

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

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

Shaklee, H., & Mims, M. (1982). Source of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Motivation, and Cognition, 8*, 208-224.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory and Cognition, 8*, 359-467.

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

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

Shanks, D. R. (1993). Human instrumental learning: A critical review of data and theory. *British Journal of Psychology, 84*, 319-354.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), *The psychology of learning and motivation: Vol. 21. Advances in research and theory* (pp. 229-261). San Diego, CA: Academic Press.

Shanks, D. R., & Lopez, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory and Cognition, 24*, 511-522.

Sidman, M. (1960). *Tactics of scientific research.* Boston: Authors Cooperative, Inc.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*, 166-173.

Tuckey, M. R., & Brewer, N. (2003). The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied*, 9, 101-118.

Von Hecker, Ulrich. (2004). Disambiguating a mental model: Influence of social context. *Psychological Record, 54*, 21-43.

Waldmann, M. R., & Holyoak, K. J. (1997). Determining whether causal order affects cue selection in human contingency learning: Comments on Shanks and Lopez (1996). *Memory and Cognition, 25*, 126-134.

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

Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the environment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 27-82). San Diego, CA: Academic Press.

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

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 509-521.

Wasserman, E. A., Kao, S., Van Hamme, L. J., Katagari, M., & Young, M. E. (1996). Causation and association. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), Causal learning (pp. 207-264). San Diego, CA: Academic Press.

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

Wasserman, E. A., & Shaklee, H. (1984). Judging response-outcome relations: The role of response-outcome contingency, outcome probability, and method of presentation. *Memory and Cognition, 12*, 270-286.

White, P. A. (2003). Effects of wording and stimulus format on the use of contingency information in causal judgment. *Memory and Cognition, 32*, 231-242.

Yates, J. F., & Curley, S. P. (1986). Contingency judgment: Primacy effects and attention decrement. *Acta Psychologica, 62*, 298-302.

APPENDICES

APPENDIX A Tables

APPENDIX A

Table A1.

Means and Standard Deviations of Each Judged Problem Set in Experiment 1.

Block	Causal Information	Contingency (Δp)	Judgment	
			M	*SD*
1	11a, 4b, 1c, 8d	.62	.62	.23
	4a, 1b, 3c, 16d	.64	.54	.26
	4a, 4b, 1c, 15d	.44	.54	.27
	2a, 12b, 0c, 10d	.14	.50	.29
2	6a, 6b, 6c, 6d	.00	.52	.26
	4a, 4b, 8c, 8d	.00	.46	.21
	8a, 5b, 9c, 2d	-.20	.53	.24
	1a, 5b, 3c, 15d	.00	.51	.30
3	1a, 8b, 11c, 4d	-.62	.51	.26
	4a, 11b, 8c, 1a	-.62	.51	.21
	4a, 4b, 15c, 1d	-.44	.51	.25
	12a, 2b, 10c, 0d	-.14	.60	.23
CS	6a, 6b, 6c, 6d	.00	.54	.27

Note. Judgments of strength were divided by 100 to facilitate comparisons with contingency. Δp = delta p; *SD* = standard deviations; CS = comparison set.

APPENDIX A continued

Table A2.
Number of Participants in Each Experimental Condition in Experiment 1.

Slide Duration	Block Order	Type of Cue	
		Abstract	Context
3 sec	1-2-3	7	11
	2-3-1	11	12
	3-1-2	11	9
5 sec	1-2-3	11	10
	2-3-1	10	10
	3-1-2	8	10

APPENDIX A continued

Table A3.

Means and Standard Deviations of Absolute and Signed Cell Deviations in Experiment 1.

Slide Duration	Contingency Table Cell				
	a	b	c	d	All Cells
	Mean Absolute Cell Deviations				
3 Second	0.95 (0.83)	1.15 (0.79)	1.20 (1.01)	1.10 (0.98)	1.10 (0.05)
5 Second	0.60 (0.65)	0.65 (0.60)	0.59 (0.55)	0.65 (0.62)	0.62 (0.02)
All Conditions	0.78 (0.76)	0.90 (0.74)	0.90 (0.87)	0.88 (0.85)	0.87 (0.03)
	Mean Signed Cell Deviations				
3 Second	.21 (.73)	-.17 (.53)	-.15 (.48)	-.03 (.64)	-.04 (.09)
5 Second	.19 (.48)	-.09 (.43)	.01 (.42)	-.01 (.46)	.03 (.06)
All Conditions	.20 (.61)	-.13 (.48)	-.07 (.45)	-.02 (.56)	-.03 (.08)

Note. Numbers in parentheses are standard deviations (for cells a, b, c, and d) and the standard error of the means (for All Cells).

APPENDIX A continued

Table A4.

Percentage of Recalled Frequencies that were Overt, Correct, and Under Estimations of the Contingency Table Cells in Experiment 1.

Type of Estimation	Contingency Table Cell			
	a	b	c	d
	All Conditions (*n* = 1433)			
% Over	21.4	19.2	20.0	21.4
% Correct	61.9	56.5	58.3	60.0
% Under	16.7	24.4	21.8	18.6
Mean	.05 (.62)	-.05 (.66)	-.02 (.65)	-.03 (.63)
	3 S Slide Duration Only (*n* = 731)			
% Over	25.7	23.3	23.9	25.4
% Correct	54.6	46.9	48.4	51.8
% Under	19.7	29.8	27.6	22.7
Mean	.06 (.67)	-.07 (.73)	-.04 (.72)	-.03 (.69)
	5 S Slide Duration Only (*n* = 702)			
% Over	17.0	15.0	15.8	17.1
% Correct	69.5	66.4	68.5	68.5
% Under	13.5	18.7	15.7	14.4
Mean	.03 (.55)	-.04 (.58)	.00 (.56)	.03 (.56)

Note. Numbers in parentheses are standard deviations.

APPENDIX A continued

Table A5.

Percentage of Participants Using Various Rule-Based Strategies in Experiment 1.

	Rule-Based Strategy					
Condition	Zero	Cell A	A Vs B	dD	dP	Unclassifiable
			Percentage of Participants			
3 s ($n = 61$)	14.8	27.9	8.2	11.5	0.0	37.7
5 s ($n = 59$)	10.2	13.6	13.6	28.8	1.7	32.2
All ($n = 120$)	12.5	20.8	10.8	20.0	0.8	35.0

Note. Percentages are based on the total number of participants in their respective groups; therefore, percentages only sum to 100 percent when added across individual rows. Strategies used by participants were determined by the Rule-Analysis Technique.

Table A6.
Intercorrelations Between the Absolute Values of Judged Strength, Judged Contingency, and Inferred Contingency in Experiment 1.

	Judgment	Contingency	Inferred Contingency
Judgment	--	[a].057*	[b].055**
Contingency		--	[c].832**
Inferred Contingency			--

[a]n = 1420. [b]n = 1415. [c]n = 1431.
*$p < .05$. **$p < .01$.

Table A7.
Descriptive Statistics for the Sum of Deviations for Each Problem Set in Experiment 1.

Problem Set	Range of Sums		% Correct Sum (Sum of Events)	Mean Sum of Deviations		
	Min	Max		Σfs	ΣDeviations	*SD*
1	15	38	74.8 (24)	23.95	-.05	2.06
2	19	28	79.2 (24)	23.87	-.13	1.03
3	19	27	79.2 (24)	23.86	-.14	0.88
4	19	30	83.3 (24)	24.00	.00	1.10
5	17	30	77.8 (24)	23.88	-.12	1.41
6	17	50	80.7 (24)	24.07	.07	2.85
7	20	30	77.5 (24)	24.10	.10	1.12
8	19	29	86.7 (24)	23.91	-.09	1.01
9	20	40	83.8 (24)	24.25	.25	1.87
10	19	33	87.5 (24)	24.03	.03	1.18
11	19	26	84.2 (24)	23.86	-.14	0.85
12	17	28	88.3 (24)	23.93	-.08	1.09
CS	20	32	84.8 (24)	24.07	.07	1.17

Note. Numbers in parentheses are the sum of event states that define each Problem Set. CS = comparison set.

Table A8.
Means and Standard Deviations of Each Judged Problem Set in Experiment 2.

Problem Set	Type of Cue	Causal Information		Δ*p*	Judgment	
		Cell Frequencies	Σ		*M*	*SD*
1	Context	1a, 8b, 12c, 4d	25	-.64	-.17	.52
2	Abstract	4a, 4b, 9c, 9d	26	.00	-.30	.33
3	Context	4a, 4b, 1c, 14d	23	.43	-.22	.43
4	Abstract	11a, 2b, 9c, 0d	22	-.78	.52	.30
5	Context	6a, 6b, 6c, 6d	24	.00	.04	.23
6	Abstract	5a, 1b, 3c, 13d	22	.65	.13	.53
7	Context	4a, 4b, 14c, 1d	23	-.43	-.25	.48
8	Abstract	1a, 5b, 3c, 15d	24	.00	-.43	.60
9	Context	13a, 4b, 1c, 8d	26	.65	.54	.32
10	Abstract	4a, 10b, 7c, 1d	22	-.59	-.41	.30
11	Context	9a, 5b, 7c, 3d	24	-.06	.26	.32
12	Abstract	2a, 12b, 0c, 9d	23	.14	-.62	.41
CS	Context II	8a, 5b, 9c, 2d	24	-.20	.12	.31

Note. Judgments of strength were divided by 100 to facilitate comparisons with contingency. Δ*p* = delta p; *SD* = standard deviation; CS = comparison set.

Table A9.

Number of Participants in Each Experimental Condition in Experiment 2.

	Number of Events		
Set Placement	Known	Unknown	Total
Early	15	13	28
Late	14	15	29
Total	29	28	57

Table A10.
Means and Standard Deviations of Absolute and Signed Cell Deviations in Experiment 2.

Condition	Contingency Table Cell				
	a	b	c	d	All Cells
	Mean Absolute Cell Deviations				
Known	1.09 (.71)	1.26 (.69)	1.45 (.90)	1.20 (.90)	1.25 (.07)
Unknown	1.23 (.83)	1.24 (.67)	1.34 (.81)	1.27 (.68)	1.27 (.02)
All Conditions	1.16 (.76)	1.25 (.67)	1.39 (.85)	1.23 (.79)	1.26 (.05)
	Mean Signed Cell Deviations				
Known	.42 (.58)	.15 (.55)	-.28 (.52)	-.26 (.58)	.03 (.16)
Unknown	.56 (.83)	-.21 (.73)	-.32 (.61)	-.19 (.76)	-.02 (.20)
All Conditions	.48 (.71)	-.03 (.67)	-.30 (.56)	-.22 (.67)	.00 (.17)

Note. Numbers in parentheses are standard deviations (for cells a, b, c, and d) and the standard error of the means (for All Cells).

Table A11.
Percentage of Recalled Frequencies that were Over, Correct, and Under Estimations of the Contingency Table Cells in Experiment 2.

Type of Estimation	Contingency Table Cell			
	a	b	c	d
	All Conditions (n = 735)			
% Over	31.8	27.3	25.0	28.7
% Correct	48.8	43.3	45.0	45.8
% Under	19.4	29.3	29.9	25.5
Mean	.12 (.71)	-.02 (.75)	-.05 (.74)	.03 (.74)
	Known Condition Only (n = 344)			
% Over	29.3	31.2	24.5	25.9
% Correct	49.3	41.1	45.3	49.9
% Under	21.3	27.7	30.1	24.3
Mean	.08 (.71)	.03 (.77)	-.06 (.74)	.02 (.71)
	Unknown Condition Only (n = 361)			
% Over	34.3	23.3	25.6	31.6
% Correct	48.3	45.7	44.7	41.6
% Under	17.4	31.0	29.7	26.9
Mean	.17 (.70)	-.08 (.73)	-.04 (.74)	.05 (.76)

Note. Numbers in parentheses are standard deviations.

Table A12.
Descriptive Statistics for the Sum of Deviations for Each Problem Set in Experiment 2.

Problem Set	Range of Sum of Estimates		% Correct Sum	Mean Sum of Estimates		
	Min	Max		Actual	Absolute	*SD*
			Known Set Information			
1	10	25	79.3 (25)	24.91	-0.81	3.00
2	24	27	89.7 (26)	25.93	-0.07	0.46
3	23	30	79.3 (23)	23.55	0.55	1.48
4	21	23	93.1 (22)	22.00	0.00	0.27
5	22	25	93.1 (24)	23.97	-0.03	0.42
6	22	24	89.7 (22)	22.14	0.14	0.44
7	22	26	86.2 (23)	23.07	0.07	0.65
8	24	26	96.6 (24)	24.07	0.07	0.37
9	25	30	86.2 (26)	26.10	1.10	0.82
10	20	23	89.7 (22)	21.93	-0.07	0.46
11	22	24	96.6 (24)	23.93	-0.07	0.37
12	23	32	89.7 (23)	23.38	0.38	1.68
CS	24	25	86.2 (24)	24.14	0.14	0.35
			Unknown Set Information			
1	12	35	21.4 (25)	23.42	-1.58	5.20
2	22	40	14.3 (26)	26.93	0.93	4.23
3	17	25	28.6 (23)	22.50	-0.50	1.84
4	15	26	39.3 (22)	22.11	0.11	2.59
5	16	33	32.1 (24)	23.75	-0.25	4.23
6	15	28	32.1 (22)	22.61	0.61	3.04
7	17	31	25.0 (23)	23.07	0.07	2.78
8	21	30	39.3 (24)	24.22	0.22	1.97
9	19	35	25.0 (26)	25.36	0.36	3.18
10	15	27	42.9 (22)	21.57	-0.43	2.66
11	13	29	21.4 (24)	22.93	-1.07	3.51
12	13	30	46.4 (23)	23.57	0.57	2.78
CS	20	28	21.4 (24)	24.19	0.19	2.02

Note. Numbers in parentheses are the sum of events that define each problem set. CS = comparison set.

Table A13.

Percentage of Participants Using Various Rule-Based Strategies in Experiment 2.

Condition	Zero	Cell A	A Vs B	dD	dP	Unclassifiable
	Rule-Based Strategy					
	Percentage of Participants					
All (n = 57)	8.9	19.6	33.9	7.1	0.0	30.4
	Set Information					
Known[a]	10.3	17.2	34.5	6.9	0.0	31.0
Unknown[b]	7.4	22.2	33.3	7.4	0.0	29.6
	Set Placement					
Early[b]	7.4	18.5	40.7	3.7	0.0	29.6
Late[c]	10.3	20.7	27.6	10.3	0.0	31.0

Note. Percentages are based on the total number of participants in their respective groups; therefore, percentages only sum to 100 percent when added across individual rows. Strategies used by participants were determined by the Rule-Analysis Technique.

[a,b]n = 29 and 28.

Table A14.
Intercorrelations Between Judged Strength, Judged Contingency, and Inferred Contingency in Experiment 2.

	Judgment	Contingency	Inferred Contingency
	All Conditions (n = 679)		
Judgment	--	[a].255*	[b].285*
Contingency			.818*
Inferred Contingency			
	Known Only (n = 346)		
Judgment	--	[b].213*	[c].270*
Contingency		--	.807*
Inferred Contingency			--
	Unknown Only (n = 333)		
Judgment	--	[d].239*	[e].312*
Contingency		--	.834*
Inferred Contingency			--

[a,b,c,d,e]Participants' judgments and/or recall frequencies were incomplete, which resulted in 22, 5, 7, 17, or 20 data points being excluded from the r statistic, respectively.
*$p < .01$.

Table A15.
Intercorrelations Between Judged Strength, Contingency, and Inferred Contingency in Experiment 2 Listed Separately For Each Subset of the Rule-Analysis Technique.

	Judgment	Contingency	Inferred Contingency
Cell A Subset: Problem Sets 1, 5, 9 (*n* = 167)			
Judgment	--	[a].609**	[b].650**
Contingency		--	.773**
Inferred Contingency			--
A Versus B Subset: Problem Sets 2, 6, 10 (*n* =171)			
Judgment	--	[c].476**	[c].508**
Contingency		--	.923**
Inferred Contingency			--
Sum Diagonals Subset: Problem Sets 3, 7, 11 (*n* = 171)			
Judgment	--	[d]-.013	[d].184*
Contingency		--	.832**
Inferred Contingency			--
Conditional Probabilities (Δp) Subset: Problem Sets 4, 8, 12 (*n* = 170)			
Judgment	--	[e]-.696**	[a]-.359**
Contingency		--	.434**
Inferred Contingency			--

[a,b,c,d,e]Participants' judgments and/or recall frequencies were incomplete, which resulted in 6, 10, 7, 4, or 5 data points being excluded from the r statistic, respectively.
$*p < .05$; $**p < .01$.

Table A16.

Number of Participants in Each Experimental Condition in Experiment 3.

Event Order	Experimental Condition		
	Online	Post-Encode	Post-Retrieve
Forward	19	15	14
Backward	7	13	9
Total Participants	26	28	23

APPENDIX A continued

Table A17.

Means and Standard Deviations of Absolute and Signed Cell Deviations in Experiment 3.

Condition	Contingency Table Cell				
	a	b	c	d	All cells
	Mean Absolute Cell Deviations				
On-Line	.38 (0.75)	.27 (0.53)	.77 (1.45)	.54 (0.90)	.49 (0.11)
Post-Encode	.71 (1.15)	.43 (0.92)	.71 (1.30)	.89 (1.42)	.69 (0.10)
Post-Recall	.48 (0.73)	.48 (0.67)	.70 (1.26)	.52 (0.95)	.53 (0.06)
All	.53 (0.91)	.39 (0.73)	.73 (1.32)	.66 (1.13)	.58 (0.08)
	Mean Signed Cell Deviations				
On-Line	-.38 (0.75)	.04 (0.60)	.77 (1.45)	-.38 (0.98)	.12 (0.27)
Post-Encode	-.36 (1.31)	.14 (1.01)	.64 (1.34)	-.54 (1.59)	-.03 (0.27)
Post-Recall	-.48 (0.73)	.22 (0.80)	.61 (1.31)	-.26 (1.05)	.02 (0.24)
All	-.40 (0.98)	.13 (0.820	.68 (1.35)	-.40 (1.25)	.00 (0.26)

Note. Numbers in parentheses are the standard deviations (for cells a, b, c, and d) and the standard error of the mean (for All cells).

Table A18.

Percentage of Recalled Frequencies that were Over, Correct, and Under Estimations of the Contingency Table Cells in Experiment 3.

Type of Estimation	Contingency Table Cell			
	a	b	c	d
	All Conditions (*n* = 77)			
% Over	03.9	16.9	31.2	11.7
% Correct	67.5	71.4	66.2	61.0
% Under	28.6	11.7	02.6	27.3
Mean	-.25 (.52)	.05 (.54)	.29 (.51)	-.16 (.61)
	Forward Event Sequence Only (*n* = 48)			
% Over	2.1	14.6	37.5	10.4
% Correct	70.8	70.8	58.3	60.4
% Under	14.6	14.6	04.2	29.2
Mean	-.25 (.48)	.00 (.55)	.33 (.56)	-.19 (.61)
	Backward Event Sequence Only (*n* = 29)			
% Over	6.9	20.7	20.7	13.8
% Correct	62.1	72.4	79.3	62.1
% Under	31.0	6.9	0.0	24.1
Mean	-.24 (.58)	.14 (.52)	.21 (.41)	-.10 (.62)

Note. Numbers in parentheses are standard deviations.

Table A19.
Percentage of Recalled Frequencies that were Over, Correct, and Under Estimations of the Contingency Table Cells for Each Judgment Task Condition in Experiment 3.

Type of Estimation	Contingency Table Cell			
	a	b	c	d
	On-Line Judgment ($n = 26$)			
% Over	0.0	11.5	34.6	7.7
% Correct	73.1	76.9	65.4	61.5
% Under	26.9	11.5	0.0	30.8
Mean	-.27 (.45)	.00 (.49)	.35 (.49)	-.23 (.59)
	Memory Based – Post Encode Condition Only ($n = 28$)			
% Over	10.7	14.3	28.6	14.3
% Correct	64.3	75.0	67.9	57.1
% Under	25.0	10.7	3.6	28.6
Mean	-.14 (.59)	.04 (.51)	.25 (.52)	-.14 (.65)
	Memory Based – Post Recall Condition Only ($n = 23$)			
% Over	0.0	26.1	30.4	13.0
% Correct	65.2	60.9	65.2	65.2
% Under	34.8	13.0	4.3	21.7
Mean	-.35 (.49)	.13 (.63)	.26 (.54)	-.09 (.60)

Note. Numbers in parentheses are standard deviations.

Table A20.

Frequency (f) and Cumulative Percentage (Cum %) of Participants that Selected Each Alternatives of the Judgment of Relative Likelihood Scale in Experiment 3.

Rating	*F*	Cum %
3	3	4.0
2	24	36.0
1	15	56.0
0	4	61.3
-1	15	81.3
-2	13	98.6
-3	1	100.0

n = 75[b]

[b]Two subjects did not complete the relative likelihood scale and were not included in this analysis.

Table A21.

Intercorrelations Between Inferred Contingency, Judgments of Relative Likelihood, Line Judgments, and Numeric Estimates of Contingency in Experiment 3 Listed Separately For Judgment Tasks Condition.

	Relative Likelihood	Line Judgment	Numeric Estimate
All Conditions			
Relative	--	.620	.600
Line Judgment		--	.963
Numeric Estimate			--
On-Line Condition			
Relative	--	.631	.621
Line Judgment		--	.952
Numeric Estimate			--
Memory-Based Post Encode Condition			
Relative	--	.479	.516
Line Judgment		--	.966
Numeric Estimate			--
Memory-Based Post Recall Condition			
Relative	--	.781	.669
Line Judgment		--	.977
Numeric Estimate			--

Table A22.

Participants' Judgments of Relative Likelihood, Line Judgments, and Numeric Estimates of Contingency That Had at Least One Discrepancy in Judgment Accuracy.

Subject ID	Relative Likelihood	Line Judgment	Numeric Estimate
2202	-1[a]	0.96	0.98
5885	-2[a]	0.51	0.5
4804	-1[a]	0.16	-0.1[a]
0397	-2[a]	0.56	0.6
6203	-1[a]	0.54	0.45
0169	-2[a]	0.81	0.8
7206	-1[a]	0.52	--
4918	-2[a]	0.68	0.65
2004	-2[a]	0.29	0.25
8005	-1[a]	0.6	0.62
6995	-1[a]	0.37	0.3
5694	-2[a]	0.75	0.8
8400	2	--	-0.2[a]

Note. Judgments are incorrect in that they were perceived as noncontingent or negative.

Table A23.
Mean Absolute Cell Deviations Calculated Separately For Each Problem Set in Experiment 2.

PS	Causal Information	Contingency Table Cell			
		a	b	c	d
1	1a, 8b, 12c, 4d	3.35 (3.60)	2.15 (2.03)	3.25 (3.36)	1.66 (1.33)
2	4a, 4b, 9c, 9d	1.32 (1.45)	1.53 (1.69)	1.26 (1.34)	1.44 (1.41)
3[a]	4a, 4b, 1c, 14d	0.74 (0.88)	1.26 (1.38)	1.30 (2.05)	1.47 (1.96)
4[a]	11a, 2b, 9c, 0d	0.63 (1.11)	1.21 (1.74)	1.72 (2.22)	0.67 (1.90)
5	6a, 6b, 6c, 6d	1.16 (1.36)	1.39 (1.60)	1.23 (1.28)	1.53 (1.62)
6	5a, 1b, 3c, 13d	1.02 (1.53)	0.67 (1.55)[c]	0.70 (1.15)[b]	1.60 (2.50)
7[a]	4a, 4b, 14c, 1d	0.98 (1.73)	1.19 (1.37)	1.74 (2.85)	1.04 (1.87)
8[a]	1a, 5b, 3c, 15d	0.40 (1.13)[b]	0.51 (0.76)	0.66 (0.86)	0.88 (1.73)
9	13a, 4b, 1c, 8d	1.12 (1.86)	1.11 (1.19)	1.49 (1.71)	1.77 (2.03)
10	4a, 10b, 7c, 1d	0.63 (1.01)[c]	0.95 (1.48)	0.82 (1.03)	0.47 (1.20)[b]
11	9a, 5b, 7c, 3d	1.37 (1.33)	1.30 (1.18)[c]	1.32 (1.36)	0.93 (1.05)[b]
12	2a, 12b, 0c, 9d	1.35 (2.70)	1.75 (2.73)	1.30 (1.99)[b]	1.33 (1.98)
Mean	--	1.17 (0.22)	1.25 (0.13)	1.40 (0.20)	1.23 (0.12)

Note. Numbers in parentheses are standard deviations (Problem Sets 1-12) and standard errors of the mean (Mean). The means across all problem sets were calculated from the numbers included in this table and differ from that previously reported due to rounding.

Table A24.
Mean Signed Cell Deviations Calculated Separately For Each Problem Set in Experiment 2.

PS	Causal Information	Contingency Table Cell			
		a	b	c	d
1	1a, 8b, 12c, 4d	3.28 (3.67)	-1.58 (2.51)	-3.09 (3.50)	0.15 (2.13)
2	4a, 4b, 9c, 9d	0.83 (1.78)	1.11 (2.00)	-0.77 (1.68)	-0.74 (1.89)
3	4a, 4b, 1c, 14d	-0.35 (1.09)[a]	0.11 (1.88)	1.23 (2.10)	-0.95 (2.26)
4	11a, 2b, 9c, 0d	0.04 (1.28)[a]	0.96 (1.89)	-1.61 (2.30)	0.67 (1.90)
5	6a, 6b, 6c, 6d	0.49 (1.72)	0.23 (2.11)	-0.63 (1.67)[a]	-0.23 (2.22)[a]
6	5a, 1b, 3c, 13d	0.63 (1.73)	0.63 (1.57)	0.32 (1.31)	-1.21 (2.71)
7	4a, 4b, 14c, 1d	0.42 (1.95)	0.07 (1.82)	-1.28 (3.09)	0.86 (1.96)
8	1a, 5b, 3c, 15d	0.37 (1.14)	-0.16 (0.90)[a]	0.41 (1.01)	-0.49 (1.88)
9	13a, 4b, 1c, 8d	-0.70 (2.06)	0.26 (1.61)	1.35 (1.83)	-1.18 (2.43)
10	4a, 10b, 7c, 1d	0.18 (1.18)	-0.67 (1.63)	-0.19 (1.30)	0.44 (1.21)
11	9a, 5b, 7c, 3d	-0.25 (1.90)	0.14 (1.76)	-0.75 (1.75)	0.30 (1.38)
12	2a, 12b, 0c, 9d	1.04 (2.83)	-1.54 (2.85)	1.30 (1.99)	-0.32 (2.37)
Mean	--	0.44 (0.31)	-0.04 (0.25)	-0.31 (0.38)	-0.23 (0.21)

Note. Numbers in parentheses are standard deviations (Problem Sets 1-12) and standard errors of the mean (Mean). The means across all problem sets were calculated from the numbers included in this table and differ from that previously reported due to rounding.

Table A25.

Mean Judgments and Inferred Contingency (Δp') of Each Judged Problem Set in Experiment 4.

PS	Condition	*f*	Δ*p*	Δ*p'*		Judgment	
				M	*SD*	*M*	*SD*
1		4a, 4b, 9c, 9d					
	On-Line		.00	--	--	-.11	.41
	Memory Based		.00	.08	.12	--	--
	All conditions		.00	--	--	--	--
2		5a, 5b, 13c, 1d					
	On-Line		-.43	--	--	.03	.56
	Memory Based		-.43	-.25	.20	--	--
	All conditions		-.43	--	--	--	--
3		9a, 5b, 7c, 4d					
	On-Line		.01	--	--	.11	.41
	Memory Based		.01	.04	.09	--	--
	All conditions		.01	--	--	--	--
4		9a, 3b, 3c, 8d					
	On-Line		.47	.20	.21	.21	.36
	Memory Based		.47	.14	.23	.27	.26
	All conditions		.47	.17	.22	.24	.32
5		4a, 10b, 7c, 4d					
	On-Line		-.35	-.21	.16	-.27	.36
	Memory Based		-.35	-.21	.23	-.28	.34
	All conditions		-.35	-.21	.20	-.28	.35

Note. Numbers in parentheses are standard deviations.

Table A26.
Means and Standard Deviations of Absolute Cell Deviations in Experiment 4.

PS Condition	Contingency Table Cell a	b	c	d	All cells
	Mean Absolute Cell Deviations				
1					
Memory-Based	2.11 (1.99)	1.75 (1.66)	2.58 (2.14)	1.68 (1.59)	2.03 (0.21)
2					
Memory-Based	1.90 (2.29)[b]	1.42 (1.46)[b]	3.15 (3.36)[d]	1.34 (1.76)[a]	1.96 (0.42)
3					
Memory-Based	1.27 (1.48)	0.56 (0.85)	0.83 (0.93)	1.00 (1.50)	0.92 (0.15)
4					
On-Line	2.85 (1.96)	1.17 (1.28)[a]	1.29 (1.18)[b]	3.23 (2.03)	2.14 (0.53)
Memory-Based	2.19 (1.72)[c]	2.06 (1.76)[b]	2.04 (1.57)[a]	2.63 (1.82)[d]	2.23 (0.14)
All Conditions	2.52 (1.86)[c]	1.61 (1.59)[a]	1.67 (1.43)[b]	2.94 (1.94)[d]	2.19 (0.33)
5					
On-Line	1.40 (1.61)	1.96 (2.29)	1.15 (1.17)	0.92 (1.03)[a]	1.36 (0.22)
Memory-Based	1.65 (1.66)	2.38 (2.30)	1.10 (1.26)	1.19 (1.78)	1.58 (0.29)
All Conditions	1.52 (1.63)	2.17 (2.29)	1.13 (1.21)	1.05 (1.45)[a]	1.47 (0.26)
--					
On-Line	1.82 (1.86)[c]	1.63 (1.77)[b]	1.94 (2.20)[d]	1.56 (1.78)[a]	1.74 (0.09)
Memory-Based	2.12 (1.92)	1.56 (1.89)	1.22 (1.17)	2.07 (1.98)	1.74 (0.22)
All Conditions	1.90 (1.88)	1.61 (1.80)	1.74 (1.99)	1.71 (1.85)	1.74 (0.06)

Note. Numbers in parentheses are the standard deviations (for cells a, b, c, and d) and the standard error of the mean (for All Cells).

Table A27.
Means and Standard Deviations of Signed Cell Deviations in Experiment 4.

PS Condition	Contingency Table Cell a	b	c	d	All cells
	Mean Signed Cell Deviations				
1					
Memory-Based	1.68 (2.37)	1.29 (2.04)	-2.12 (2.61)	-0.28 (2.31)	0.14 (0.86)
2					
Memory-Based	1.23 (2.72)[a]	-0.12 (2.04)[a]	-2.98 (3.51)	1.30 (1.79)	-0.14 (1.00)
3					
Memory-Based	-0.56 (1.88)	0.19 (1.00)[a]	-0.63 (1.08)	0.79 (1.62)	-0.05 (0.34)
4					
On-Line	-2.47 (2.43)	0.92 (1.47)	0.50 (1.69)	-3.15 (2.16)	-1.05 (1.03)
Memory-Based	-1.81 (2.13)	1.89 (1.95)	1.75 (1.90)	-2.02 (2.50)	-0.05 (1.08)
All Conditions	-2.14 (2.29)	1.40 (1.78)	1.13 (1.89)	-2.60 (2.38)	-0.55 (1.06)
5					
On-Line	1.19 (1.77)	-1.83 (2.40)	-0.40 (1.60)	0.38 (1.33)	-0.17 (0.64)
Memory-Based	1.19 (2.02)	-2.13 (2.54)	-0.27 (1.66)	0.40 (2.11)	-0.20 (0.71)
All Conditions	1.19 (1.89)	-1.98 (2.46)	-0.33 (1.62)	0.39 (1.76)	-0.18 (0.67)
--					
On-Line	-0.62 (2.80)	-0.46 (2.41)	0.05 (1.69)	-1.39 (2.52)	-0.61 (0.30)
Memory-Based	0.35 (2.58)	0.22 (2.40)	-0.85 (2.81)	0.05 (2.36)	-0.06 (0.27)
All Conditions	0.07 (2.68)	0.02 (2.42)	-0.59 (2.57)	-0.36 (2.49)	-0.22 (0.16)

Note. Numbers in parentheses are the standard deviations (for cells a, b, c, and d) and the standard error of the mean (for All Cells).

Table A28.

Percentage of Recalled Frequencies that were Correct Estimations of the Contingency Table Cells for Problem Sets Four and Five of Experiment 4.

PS	Statistic	Contingency Table Cell			
		a	b	c	d
1		Memory-Based Condition (*n* = 48)			
	% Correct	17.0	22.9[a]	16.7[a]	21.3
	Mean	.49 (.78)	.40 (.79)	-.50 (.77)	.02 (.90)
2					
	% Correct	33.3	29.2	27.1	46.8[a]
	Mean	.25 (.79)	-.04 (.85)	-.65 (.57)	.49 (.55)
3					
	% Correct	41.7	60.4	43.8	54.2
	Mean	-.17 (.75)	.06 (.63)	-.40 (.64)	.25 (.64)
4					
	% Correct	21.3[a]	21.3[a]	18.8	17.4[b]
	Mean	-.57 (.68)	.66 (.60)	.65 (.64)	-.48 (.78)
5					
	% Correct	35.4	29.2	43.8	37.5
	Mean	.35 (.73)	-.54 (.65)	-.15 (.74)	-.08 (.80)
4		On-Line Condition (*n* =48)			
	% Correct	12.8	33.3	29.2	8.3
	Mean	-.70 (.62)	.42 (.71)	.29 (.80)	-.88 (.39)
5					
	% Correct	41.7	41.7	33.3	41.7
	Mean	.42 (.65)	-.46 (.62)	-.08 (.82)	.13 (.76)

Note. Numbers in parentheses are standard deviations.

[a,b]Recalled frequencies were missing and one or two estimations were not included in this analysis, respectively.

Table A29.
Intercorrelations Between Judged Strength, Judged Contingency, and Inferred Contingency in Experiment 4.

	Judgment	Contingency	Inferred Contingency
All Conditions ($n = 330$)			
Judgment	--	[a].374*	[b].581*
Contingency		--	.661*
Inferred Contingency			
On-Line Condition ($n = 232$)			
Judgment	--	.247*	[c].616*
Contingency		--	[d].734*
Inferred Contingency			--
Memory Based Condition ($n = 235$)			
Judgment	--	[e].674*	[c].550*
Contingency		--	.618*
Inferred Contingency			--
On-Line Condition (Problem Sets 4 and 5; $n = 95$)			
Judgment	--	[f].560*	[g].616*
Contingency		--	.734*
Inferred Contingency			--

[a,b,c,d,e,f,g]Participants' judgments and/or recall frequencies were incomplete, which resulted in 3, 147, 142, 137, 140, 4, or 5 data points being excluded from the r statistic, respectively.
*$p < .01$.

APPENDIX B Figures

APPENDIX B

	OUTPUT	NO OUTPUT
INPUT	a	b
NO INPUT	c	d

Figure B1. A summary table depicting the four event states, their labels, and the unique information about the input and output that each cell represents. This table is typically referred to as a contingency table.

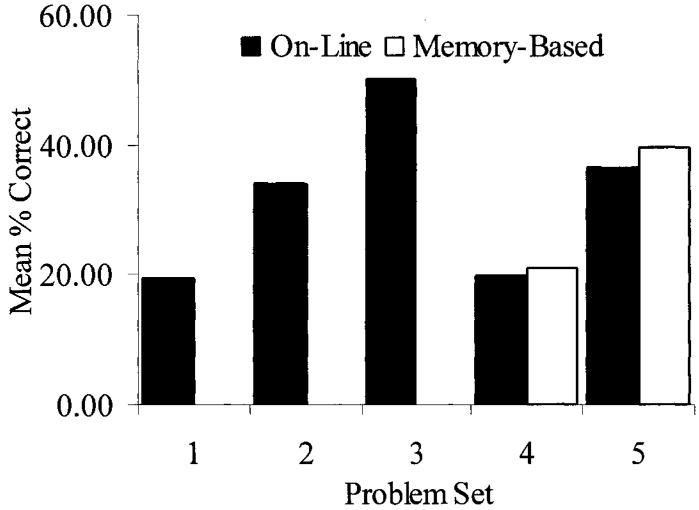

Figure B2. The mean percentages of correct cell estimations across each problem set in Experiment 4. Means were calculated across all four cells of the problem set. Empty and black bars represent the percentage of correct estimations obtained by participants in the on-line and memory-based conditions, respectively.

APPENDIX C Evidence of IRB Approval Forms

APPENDIX C

University of New Hampshire
Institutional Review Board for the Protection of Human Subjects in Research
Departmental Review Committee Exemption Classification Sheet

Project Director Christopher Barnes **IRB #** 11
Department Psychology **Reviewer** ____
Project Title Effects of Memory for Serially Presented Causal Cues

Reviewer: Please write comments or contingencies of approval, if any, on a separate sheet of paper, and attach to this form. Place the completed form on file with the application for review, in the Departmental Review Committee files. Protocol applications and review forms will be forwarded to the Office of Sponsored Research each semester for reporting purposes.

☑ **Protocol qualifies as EXEMPT under the following subsection (check one) - see reverse for detailed category description:**

- ✓ 46.101(b)(1) Research conducted in established educational setting using normal educational procedures
- ___ 46.101(b)(2) Educational tests, surveys, interviews, observation of public behavior/no risk
- ___ 46.101(b)(3) Educational tests, surveys, interviews, observation of public behavior not exempt under Subsection 2, above, if public official or if confidentiality mandated by federal statutes
- ___ 46.101(b)(4) Study of existing data
- ___ 46.101(b)(5) Study of public benefits or service programs
- ___ 46.101(b)(6) Taste and food studies

☐ **Refer protocol to the regular IRB for EXPEDITED review under the following subsection (check one):**

- ___ 46.110(b)(1) Clinical studies of drugs/medical devices not requiring investigational new drug/device applications.
- ___ 46.110(b)(2) Collection of blood samples by finger, heel or ear stick, or venipuncture in healthy adults >110 lbs., or others and children, considering age, weight, health, collection procedure, frequency and amount of collection.
- ___ 46.110(b)(3) Prospective collection of biological specimens for research purposes by noninvasive means, and in a non-disfiguring manner: hair and nail clippings, teeth, sweat, saliva, placenta (after delivery), amniotic fluid (at membrane rupture/labor), dental plaque/calculus, mucosal/skin cells, sputum (after saline nebulization)
- ___ 46.110(b)(4) Collection of data through noninvasive means routinely employed in clinical practice (excluding x-rays and microwaves, and devices not approved for marketing): physical sensors applied to the skin, weighing, tests of visual acuity, MRI, EKG, EEG, ultrasound, etc., and moderate exercise by healthy volunteers.
- ___ 46.110(b)(5) Non-exempt research involving data, documents, records or specimens that have been/will be collected solely for nonresearch purposes (e.g., medical treatment or diagnosis).
- ___ 46.110(b)(6) Collection of data from voice, video, digital, or image recordings made for research purposes.
- ___ 46.110(b)(7) Non-exempt research on individual or group behavior or characteristics of individuals, such as studies of perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior, or research employing surveys, interviews, oral histories, focus groups, program evaluation, human factors evaluation, or quality assurance methodologies.
- ___ 46.110(b)(8) Continuing review of research such as studies permanently closed to enrollment of new subjects, or for which research-related interventions are completed, or for which only long-term follow-up of subjects remains, or for which no subjects have been enrolled and no additional risks have been identified, or for which data analysis is the only remaining research activity.
- ___ 46.110(b)(9) Continuing review of research (not conducted under investigational drug/device applications or exemption) where categories 2 through 8, above, do not apply, and for which the IRB has determined that the research involves no greater than minimal risk, and no additional risks have been identified.

☐ **Refer protocol to the regular IRB for FULL BOARD action (cite reason on separate sheet)**

☐ **Protocol cannot be approved as presented (cite reason on separate sheet)**

IRB Reviewer: [signature] Date: 7/22/04

University of New Hampshire
Institutional Review Board for the Protection of Human Subjects in Research
Departmental Review Committee Exemption Classification Sheet

Name: Christopher Barnes **IRB #:** 10

Dept: Psychology **Reviewer:** ______

Study: Effect of Memory for Serially Presented Causal Cues

Exempt Review

46.101(b)(1) ______ Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as:
(i) research on regular or special educational instructional strategies, or
(ii) research on the effectiveness of or comparison among instructional techniques, curricula, or classroom management methods.

46.101(b)(2) ✓ Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior unless:
(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and
(ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to subjects' financial standing, employability, or reputation.

46.101(b)(3) ______ Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior that is not exempt under category (b)(2) if:
(i) the human subjects are elected or appointed public officials or candidates for public office; or
(ii) federal statute(s) require(s) without exception that confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.

46.101(b)(4) ______ Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

46.101(b)(5) ______ Research and demonstration projects which are conducted by or subject to the approval of department or agency heads, and which are designed to study, evaluate, or otherwise examine: (i) public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs.

46.101(b)(6) ______ Taste and food quality evaluation and consumer acceptance studies, (i) if wholesome foods without additives are consumed or (ii) or if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration, or approved by the Environmental Protection Agency, or the Food Safety and Inspection Service of the U.S. Department of Agriculture.

______ **Protocol is approved as presented in the category checked**
______ **Protocol is approved with the following contingencies/comments (attach sheets if necessary)**
______ **Protocol is referred to the IRB for Expedited or Full Board review**
______ **Protocol cannot be approved as presented (cite reasons on separate sheet)**

DRC Reviewer: [signature] **Date:** 9/9/04

University of New Hampshire
Institutional Review Board for the Protection of Human Subjects in Research
Departmental Review Committee Exemption Classification Sheet

Name: Christopher Barnes **IRB #:** 28
Dept: Psychology **Reviewer:** ______
Study: Effects of Memory for Serially Presented Causal Cues

Exempt Review

46.101(b)(1) ______ Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as:
(i) research on regular or special educational instructional strategies, or
(ii) research on the effectiveness of or comparison among instructional techniques, curricula, or classroom management methods.

46.101(b)(2) ✓ Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior unless:
(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and
(ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to subjects' financial standing, employability, or reputation.

46.101(b)(3) ______ Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior that is not exempt under category (b)(2) if:
(i) the human subjects are elected or appointed public officials or candidates for public office; or
(ii) federal statute(s) require(s) without exception that confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.

46.101(b)(4) ______ Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

46.101(b)(5) ______ Research and demonstration projects which are conducted by or subject to the approval of department or agency heads, and which are designed to study, evaluate, or otherwise examine: (i) public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs.

46.101(b)(6) ______ Taste and food quality evaluation and consumer acceptance studies, (i) if wholesome foods without additives are consumed or (ii) or if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration, or approved by the Environmental Protection Agency, or the Food Safety and Inspection Service of the U.S. Department of Agriculture.

______ **Protocol is approved as presented in the category checked**
______ **Protocol is approved with the following contingencies/comments (attach sheets if necessary)**
______ **Protocol is referred to the IRB for Expedited or Full Board review**
______ **Protocol cannot be approved as presented (cite reasons on separate sheet)**

DRC Reviewer: [signature] **Date:** 3/9/05

APPENDIX D Laboratory Experience Information Forms

APPENDIX D

LABORATORY EXPERIENCE INFORMATION FORM

Experiment Number: 11

Experimenter: **Christopher Alexander Barnes**

General Area of Psychology:
Social - Cognitive Psychology: More specifically this research falls under human judgment of covariation.

General problem or issue under investigastion:
This experiment aims to isolate whether the tendency to make differential errors in the recreation of a summary table is the result of biases at the encoding or retrieval portion of the memory process. The main independent varable is the point at which subjects are informed of the intent to form a judgment of the relation between the two variables.

Specific hypothesis:
1 - Subjects with informed knowledge of the intention to form a judgment prior to viewing the slides will differentially make errors.
2 - The bias of differentially making errors stems from an encoding error influenced by the intent to form a judgment prior to vieweing the causal information.

Correlational or independent and dependent variables:
Independent variable: Point at which subjects of informed of the task to form a judgment.
1 - Normal (prior to viewing) 2 - Encode-Free (after viewing) 3 - Retreival-Free (after task) 4 - Never
Dependent variable: The absolute value of the deviations from perceived and actual cell frequencie

Control procedures:
To assess whether forming a judgment biases the memory of causal cues, the control group (Never) will never be introduced to the bias of forming a judgment. The Normal group will be introduced to the bias in a manner consistent with that traditionally in the literature.

Potential implications of study:
This study plans to demonstrate that a subject's knowledge of forming a judgment may serve as a potentially biasing agent in either the encoding or retrieval of events.

LABORATORY EXPERIENCE INFORMATION FORM

Experiment Number: 77

Experimenter: Christopher Alexander Barnes

General Area of Psychology:
Social - Cognitive Psychology: More specifically this research falls under human judgment of covariation.

General problem or issue under investigastion:
The accuracy in which indiviuals are able to recalled causally relevant information from memory when the information is presented to them in a serial fashion. It is expected that recalled events will be inaccurate and this may serve as a potential contributor to erroneous judgments of contingency.

Specific hypothesis:
1 - Subjects that have a longer exposure to the co-events will have better accuracy in their recalled estimates of the actual co-events presented.
2 - Subjects that are presened the co-events in a plausible context will have produce more errors in their recalled of the actual co-events.

Correlational or independent and dependent variables:
Independent variables: Exposure time to co-events (3s vs, 5s)
Context of the co-events (plausibly relevents events vs. abstract events)
Dependent variable: The absolute value of the deviations from perceived and actual frequencies.

Control procedures:
To assess whether previous attempts to produce differential errors in recalled error is a by-product of the extended amout of exposure time in previous work, the current study will expose subjects to a condition which will increase the difficulty of the task.

Potential implications of study:
This study plans to demonstrate that under circumstance when strategies can not be used to encode and recall causal information that causal information is recalled in abiased manner. To assess whether the the context of the events influences recall.

LABORATORY EXPERIENCE INFORMATION FORM

Experiment Number: 25

Experimenter: **Christopher Alexander Barnes**

General Area of Psychology:
Social - Cognitive Psychology: More specifically this research falls under human judgment of covariation.

General problem or issue under investigastion:
To determine whether memory-based judgments of contingency are more highly correlated with the events recalled from memory than on-line judgments of contingency. Memory-based judgments refer to judgments that must be made after viewing the relevant information, while on-line judgments refer to those that are assessed while viewing the events.

Specific hypothesis:
1 - Participants' judgments of contingency on the memory-based task will be more highly correlated with statistically inferred based on the information recalled from memory than participant's judgments on the on-line task.
2 - In the fourth problem Set, the accuracy of recall will be better for those participants in the memory-based condition than those in the on-line condition.

Correlational or independent and dependent variables:
Independent variable: Type of Judgmnent: Memory-Based vs. On-Line

Dependent variable: Absolute value of the deviations from perceived and actual frequencies; Judgments of contingency.

Control procedures:
The two Type of Judgment conditions serve as the control conditions. Forewarning participants of the judgment task (On-Line Condition) is expected to result in more accuracte judgments of contingency, while not forewarning participants (Memory-Based) is expected to result in judgments that are more consistent with what they recall from memory—not what was shown. Therefore, predicted differences between conditions will serve to prove an expermental effect.

Potential implications of study:
This study will highlight the importance of recognizing task differences, on-line vs. memory-based tasks, in the contingency judgment literature. Real world judgments tend to be memory-based, yet experiments tend to use online procedures to investigate contingency judgment.