Integration of Contingency Information in Judgments of Cause, Covariation, and Probability

David R. Mandel
Stanford University

Darrin R. Lehman
University of British Columbia

Integration of contingency information underlies many cognitive tasks including causal, covariational, and probability judgments. The authors' *feature-analytic approach* was used to account for the findings that people differentially weight specific types of conjunctive information in causal (Experiment 1) and noncausal (Experiment 2) contingency judgments. These findings were explained in terms of positive-test and sufficiency-test biases, which were found in both judgment domains. The same biases, however, were not observed in normative conditional-probability judgments (Experiment 3). The authors argue that this discrepancy is owing to the differential clarity of normative criteria in these domains.

Much of human learning and inferential thinking depends on the integration of contingency information. To test hypotheses and revise beliefs; to explain past events and predict future ones; to establish categories, form stereotypes, and develop impressions of others, humans integrate a vast amount of information about interevent contingencies. In short, the ability to discriminate contingencies in the physical and social worlds of humans is a basic characteristic of adaptive behavior. It is important, then, to understand how people accomplish this first-order cognitive task in different domains of judgment.

Our work addresses the issue of how people go about integrating contingency information in order to arrive at intuitive statistical estimations of causality, covariation, and likelihood. The super-symbolic society (Toffler, 1990) that people now live in affords them a greater number of facts and figures to process than at any other time in history. This is an age of both statistical dependence and statistical distrust. One cannot read the newspaper, watch television, or listen to the radio without receiving a barrage of statistical information in the form of frequencies, means, probabilities, and variances. How do people use the statistical information that they receive to update their beliefs, to evaluate claims,

to form causal theories, or to make decisions? Can people decipher when others have integrated statistical information in an appropriate manner? Can people even articulate accurately how they, themselves, go about integrating statistical information in the service of judgment and decision making? The study of intuitive statistics (e.g., Busemeyer, 1991) is becoming increasingly significant, for humans are intuitive statisticians (Peterson & Beach, 1967) now more than ever before.

The specific judgment tasks examined in this article require participants to make statistical estimations based on data concerning the frequency of event conjunctions. Specifically, we have examined how people integrate (i.e., weight and combine) contingency information in three domains of judgment: causal contingency (i.e., what is the effect of X on Y?), noncausal contingency (i.e., what is the relation between X and Y?), and conditional probability (i.e., what is the likelihood of X given Y?).[1] We have examined the 2 × 2 case, shown in Figure 1, in which an input (e.g., lightning) and an output (e.g., fire) each are either present or absent (see Trolier & Hamilton, 1986, for a comparison of contingency judgments of binary vs. continuous variables). The four event conjunctions are labeled Cells A–D, and the italicized letters *A–D* denote the respective cell frequencies.

Our objectives have been threefold: First, we drew attention to a conceptual and methodological problem underlying past attempts to discriminate between various algebraic rules of information integration. Fundamentally, the problem entails an inability to attribute a rule's predictive value to specific rule features. We have proposed a new *feature-analytic approach* that addresses this problem and that uses rules to make inferences about features of human information processing. Second, we have provided a theoretical account of the robust finding that people ascribe differen-

[1] Although we focus on the integration of contingency information, it is noteworthy that contingency judgment may include other processes such as deciding what kinds of information are relevant, sampling information, classifying instances, and recalling evidence and estimating frequencies (Alloy & Tabachnik, 1984; Crocker, 1981; Fiske & Taylor, 1991; Jennings, Amabile, & Ross, 1982).

Output

	Present (O)	Absent (~O)
Present (I)	**Cell A** $A = f(I \cap O)$	**Cell B** $B = f(I \cap {\sim}O)$
Absent (~I)	**Cell C** $C = f({\sim}I \cap O)$	**Cell D** $D = f({\sim}I \cap {\sim}O)$

Input

Figure 1. The 2 × 2 contingency table: An input is either present (*I*) or absent (*~I*) and an output is either present (*O*) or absent (*~O*). Cell frequencies are denoted by the letters *A* to *D*.

tial importance to Cells A–D in contingency judgment. Although powerful new methods have been developed to study the importance that people assign to the four cells (e.g., Levin, Wasserman, & Kao, 1993; Wasserman, Dorner, & Kao, 1990), the finding that people ascribe the greatest weight to *A* followed respectively by *B*, *C*, and *D* has yet to be explained. We have accounted for this finding in terms of two biases: (a) a *positive-test* (+test) *bias* (Klayman & Ha, 1987) in which people weight information about event presence greater than information about event absence and (b) a *sufficiency-test* (Stest) *bias* in which people weight sufficiency information greater than necessity information. Third, we have examined the extent of these biases in multiple judgment domains. Whereas +test and Stest biases emerged in causal and noncausal contingency judgment, they did not emerge in normative conditional-probability judgment. We have proposed that the discrepancy stems from the fact that probability (but not contingency) judgments are guided by a transparent normative criterion. We have concluded by discussing prescriptive, normative, and additional descriptive issues.

Rule Discrimination

Often, in contingency-judgment experiments, participants first are presented with contingency information either on a case-by-case basis (e.g., Allan & Jenkins, 1980, 1983; Jenkins & Ward, 1965) or in summary form in which the frequencies for the four cells are given (e.g., Kao & Wasserman, 1993, Experiment 2; Ward & Jenkins, 1965). Participants then are asked to rate the magnitude and direction of the interevent relation. The task is repeated over a set of information patterns yielding a vector of ratings for each participant. A common data-analytic approach (e.g., Anderson & Sheu, 1995; Schustack & Sternberg, 1981) has been to regress the group-averaged ratings on the output of various algebraic rules. The objective is to find the best fitting rule.

Several rules have been examined (for recent reviews, see Allan, 1993; Kao & Wasserman, 1993).[2] One that has received considerable attention, partly because of its status as a normative model of causal contingency, is the delta rule (Jenkins & Ward, 1965; Salmon, 1965; for recent descriptive models of causal inference, see Cheng, 1997; Cheng & Holyoak, 1995),

$$\Delta P = P(O|I) - P(O|{\sim}I) = A/(A + B) - C/(C + D).$$

As shown, ΔP contrasts two conditional probabilities: the likelihood of the outcome occurring given that the input is present minus the likelihood of the outcome occurring given that the input is absent. Another rule that has received considerable attention is a weighted linear combination of the four cell frequencies (Schustack & Sternberg, 1981):

$$\text{weighted } \Delta D = \omega_A A - \omega_B B - \omega_C C + \omega_D D,$$

in which ΔD stands for the difference in the sums of the diagonals. Many other rules that have been examined are special cases of weighted ΔD. For instance, the *A*-minus-*B* rule (Inhelder & Piaget, 1958) sets $\omega_A = \omega_B = 1$ and $\omega_C = \omega_D = 0$. The confirming-cases rule (Ward & Jenkins, 1965) sets either $\omega_A = \omega_D = 1$ and $\omega_B = \omega_C = 0$ or vice versa. And the success-counting or Cell A rule (Allan & Jenkins, 1983; Nisbett & Ross, 1980; Inhelder & Piaget, 1958; Shaklee & Tucker, 1980; Smedslund, 1963) sets $\omega_A = 1$ and the remaining weights equal to zero. A joint-probability variant of weighted ΔD (i.e., weighted ΔD divided by the sum of the four frequencies) proposed by Busemeyer (1991) has been examined recently, also (Anderson & Sheu, 1995; for a variation of this rule, see Kao & Wasserman, 1993).

Literature reviews (e.g., Allan, 1993; Kao & Wasserman, 1993) indicate that although the rules examined tend to account for substantial proportions of variance in participants' ratings, no consistent rank ordering of the rules in terms of their predictive values has been found across studies. As Allan (1993, p. 438) put it, "the quest for a rule to describe human judgments of the contingency between binary variables has not yet yielded a satisfactory solution." The inconsistency has largely been attributed to experimental features. For instance, the format of information presentation is likely to influence strategy use (Beyth-Marom, 1982; Crocker, 1981; Shaklee, 1983). When information is presented in summary form, participants tend to adopt more complex strategies such as ΔP (e.g., Kao & Wasserman, 1993; Ward & Jenkins, 1965) presumably because of reductions in (a) memory demands (Kao & Wasserman, 1993; Shaklee & Mims, 1982; Yates & Curley, 1986), (b) ambiguity concerning how event pairs are defined (Wasserman & Shaklee, 1984), and (c) frequency-estimation errors (Wasserman & Shaklee, 1984). Many studies also have used

---

[2] The Rescorla–Wagner rule (Rescorla & Wagner, 1972) also has received considerable attention. This rule, however, is an acquisition model for judgments based on trial-by-trial or continuous data. Given that the reported experiments use summary information, this model is inappropriate.

leading instructions that would favor some rules over others (for reviews, see Beyth-Marom, 1982; Cheng & Novick, 1992; Crocker, 1981). In the next section, we have addressed a more general and conceptual issue about rule discrimination that bears on the question, "What would we be able to conclude even if a given rule did consistently outperform others?"

Feature Analysis

Just as any object can be described by a particular set of attributes or features, so can a given rule. There are two compatible ways to think about the relation between rules and their features: (a) features are *descriptors* of rules and (b) rules are *exemplars* of features. For instance, if it is true that Rule 1 is described by Features X, Y, and Z, then it is also true that Rule 1 is an exemplar of each of these features. It is instructive to consider the dual nature of this relation for two reasons. First, it suggests an important methodological flaw of past research. Second, and of more importance, it suggests a new method for making inferences about the features that describe human information processing.

Past attempts at rule discrimination generally are flawed because the target set of rules being evaluated simultaneously varied in terms of many different features. This makes it impossible to analyze the unique contribution of each of these features. Consequently, it is also impossible to pinpoint why a given rule may be a good predictor of human judgment. For instance, suppose we are evaluating ΔP and the A-minus-B rule, and we find that the former outperforms the latter. To what do we attribute this success? Is it because ΔP combines conditional probabilities, whereas the A-minus-B rule combines frequencies? Or is it because the former uses information from all four cells, whereas the latter only uses information from Cells A and B? Moreover, how do we compare the predictive values of these two rules given that they also differ in terms of the number of parameters they contain? These interpretational difficulties are augmented in real experiments in which, typically, four or more rules are compared. In short, past research typically has compared the predictive values of rules described by various conjunctions of features without being able to tease apart the contributions made by these features. The solution to this problem is to construct rules that systematically manipulate features while holding constant the number of parameters.

The advantages of a feature-analytic approach extend beyond simply rectifying past methodological problems. As noted earlier, the objective of rule discrimination is to identify the best predictive model from a set of candidates. In contrast, we propose that, because rules represent exemplars of features, they can be used as tools to draw inferences about features of human information processing. This is our main interest in rule use. Consider the relation between a rule and a human information processor. Both operate on information, weighting and combining it in a particular manner. The information itself may be coded in terms of features. For example, the contingency information shown in Figure 1 can be described in terms of the positivity of the

information. That is, Cell A is positive because both events are present, Cells B and C are effectively neutral because one event is present and the other is absent, and Cell D is negative because both events are absent. If people are prone to a +test bias, as we have suggested, then rules that weight positive information more heavily than negative information should correlate more strongly with people's judgments than rules that weight negative information more strongly than (or even equally to) positive information. Thus, we can systematically construct rules that differ in terms of this weighting in order to examine the importance of this feature as a descriptor of human information processing.

In order to decompose the specific contributions made by a set of target features, we introduce a variance-partitioning technique termed *feature analysis*. The logic of this technique is straightforward: If a particular feature is an important determinant of a rule's viability as a predictive model, then partitioning the variance in the correlations between rule output and human judgments in terms of levels (e.g., presence vs. absence) of that feature should yield a reliable main effect and a reasonably strong effect size. Specifically, we have defined rule viability in terms of z_{ij}, the Fisher-transformed correlation (Fisher, 1921) between the ith rule's output and the jth individual's responses.[3] In feature analysis, then, the variance in z_{ij} is partitioned in terms of a target set of features, allowing for a systematic decomposition of rule viability. In short, the objective of the feature-analytic approach is to draw inferences about features of human information processing by analyzing the predictive values of rules whose output and features both are known.

The Cell Weight Inequality

Perhaps the most consistent finding concerning how people integrate contingency information is that they assign differential importance to each of the four cell frequencies. In absolute terms (i.e., ignoring whether the value of the weight is positively or negatively valenced—the sense in which we have used the term *absolute* throughout this article), people tend to weight $A > B > C > D$ (e.g., Anderson & Sheu, 1995; Kao & Wasserman, 1993; Levin et al., 1993; Schustack & Sternberg, 1981). Wasserman et al. (1990), for example, found that this general pattern held when (a) participants were asked to indicate explicitly what types of information are important for contingency judgment (see also Crocker, 1982), (b) participants were asked to explicitly rate cell importance on a continuous scale, and (c) cell weights were determined on the basis of participants' ratings on a contingency-judgment task.

Because the preceding inequality is probabilistic, it is more accurately expressed in terms of the expected, absolute values of the cell weights:

$$\exists J[E(|\omega_A|) > E(|\omega_B|) > E(|\omega_C|) > E(|\omega_D|)]. \quad (1)$$

---

[3] In cases in which correlation coefficients are used as raw data in secondary statistical analyses, it is appropriate to first Fisher transform these data in order to increase the normality of the distribution.

Equation 1 states that there exists a set of contingency-based judgments, J, such that the expected, absolute value of the weight for A is greater than that for B and so forth. We refer to the empirical finding expressed in Equation 1 as the *cell weight inequality* (CWI). It is also consistently found that people tend to weight A and D (viz., cases confirming a positive relation) positively and they tend to weight B and C (viz., cases confirming a negative relation) negatively (for a meta-analysis demonstrating this finding, see Lipe, 1990).

Despite the consistency of the CWI, a theoretical account of why it obtains is still lacking. Such an account is required in order to explain why people's contingency judgments systematically diverge from statistical measures of contingency such as ΔP (or the phi coefficient in the case of two-way dependency; see Allan, 1980). Recently, Cheng (1997) provided a mathematical account of why $E(|\omega_A|, |\omega_B|) > E(|\omega_C|, |\omega_D|)$ in causal inference. This account does not specify why $E(|\omega_A|) > E(|\omega_B|)$, why $E(|\omega_C|) > E(|\omega_D|)$, or why the CWI would obtain in noncausal contingency judgment. The feature-analytic account that we now describe addresses each of these issues.

Positive Versus Negative Tests

Cognitive positivity bias refers to a preference for, or tendency to use, information that is stated in affirmational (positive) rather than negational (negative) terms. People exhibit several positivity biases (e.g., see Fazio, Sherman, & Herr, 1982; Halberstadt & Kareev, 1995; Klayman & Ha, 1987; McGuire & McGuire, 1992; Nisbett & Ross, 1980; also see Kareev, 1995, for a discussion of bias toward perceiving positive correlation). For instance, people tend to learn concepts more efficiently from positive information than from negative information even when the informativeness of both sources is equal (Hovland, 1952; Hovland & Weiss, 1953). They are more likely to learn to discriminate between (or predict) positive and negative states (e.g., a light being on or off) when the positive state is signaled by the presence rather than the absence of a diagnostic cue (Newman, Wolff, & Hearst, 1980). People process positive information faster than negative information (Wason, 1959). And, people tend to use affirmative statements in verbal communication (Clark & Clark, 1977).

As noted earlier, the contingency information depicted in Figure 1 can be described in terms of positivity such that Cell A is positive, Cells B and C are neutral, and Cell D is negative. In line with Klayman and Ha (1987), we define +tests as those involving a contrast between A and either B or C. Alternatively, −tests contrast D with either B or C. For example, ΔP can be described as a contrast between a +test, $P(O|I)$, and a −test, $P(O|\sim I)$. Klayman and Ha (1989) found that participants completing a rule-discovery task demonstrated a +test bias. In such tasks, a target set of events is defined by the experimenter according to a particular rule. For instance, the event may be three consecutive numbers and the rule may be "increasing numbers" (as in Wason's, 1960, classic study). Participants are given one instance of a rule-fitting case (e.g., 2, 4, 6) and are asked to determine the underlying rule by testing other instances. Klayman and Ha

found that, out of 18 hypothesis-testing opportunities, on average, participants made 12 +tests and only 2 −tests (4 tests were neither positive nor negative tests). Clearly, rule-discovery tasks can be represented in terms of 2 × 2 contingency information: A hypothesized instance can be either absent from or present in the target set, and likewise for nonhypothesized instances.

We believe that Klayman and Ha's (1989) findings are one example of a +test bias in contingency information integration and that +test bias is an important source of the CWI. A +test bias implies that

$$\exists J[E(|\omega_A|) > E(|\omega_B|) \approx E(|\omega_C|) > E(|\omega_D|)]. \quad (2)$$

As can be seen by comparing Equations 1 and 2, a +test bias would account for an important part of the CWI. Namely, both inequalities share the fact that (a) $E(|\omega_A|)$ is greater than that for the remaining weights and (b) $E(|\omega_D|)$ is less than that for the remaining weights. Supporting the notion that a +test bias underlies the CWI, the latter is attenuated when levels of the input and output factors denote the presence of two separate categories (i.e., symmetric variables; e.g., red vs. green) rather than the presence or absence of a single category (i.e., asymmetric variables; e.g., red vs. not red; see Allan & Jenkins, 1980, 1983; Beyth-Marom, 1982).

Sufficiency Versus Necessity Tests

In judging contingencies, people may ask two different types of questions: (a) "Is one event sufficient to account for, or produce, another event?" and (b) "Is one event necessary to account for, or produce, another event?" Although the relative weights attached to sufficiency and necessity criteria are almost certainly influenced by task objectives, we suspect that people display a general bias toward weighting tests of *sufficiency* (Stests) more heavily than tests of *necessity* (Ntests). Stesting has greater pragmatic value than Ntesting because, in many real-world situations, sufficiency violations may be more costly than necessity violations (Friedrich, 1993). As Klayman and Ha (1989) put it, "We don't mind passing up some potentially acceptable cars if we can avoid buying a lemon" (p. 603). Klayman and Ha's (1989) findings indicate that people tend to favor narrowing hypotheses that are too broad (i.e., insufficient hypotheses) rather than broadening hypotheses that are too narrow (i.e., unnecessary hypotheses). Similar biases toward Stesting have been observed in the domain of causal attribution (e.g., McGill & Klein, 1993, 1995). There is also evidence that counterfactual conditional thinking about negative outcomes is more indicative of thoughts about factors that might have been sufficient to prevent the outcome than factors that were necessary to cause the outcome (Mandel & Lehman, 1996).

We define sufficiency and necessity in relative terms[4]: For

---

[4] In noise-free environments in which feedback is always accurate and other factors do not influence the target relation (e.g., in rule-discovery tasks), a single case of insufficiency or nonnecessity is enough to disconfirm a sufficiency or necessity relation, respectively. In many everyday contexts, however, information

hypothesized positive (or facilitative) relations, the sufficiency of the input to account for the output increases as B decreases relative to A or D; the necessity of the input to account for the output increases as C decreases relative to A or D. For hypothesized negative (or inhibitory) relations, sufficiency increases as A decreases relative to B or C; necessity increases as D decreases relative to B or C. As can be seen, the definition of Stesting and Ntesting takes hypothesized relation valence into account. To clarify how relation valence influences such labeling, it is useful to introduce the distinction between contrasts that occur within levels of either the input (viz., contrasts of A and B or of C and D) or the output (viz., contrasts of A and C or of B and D). The former are termed *input tests* (Itests) and the latter are termed *output tests* (Otests). As shown in Figure 2, the relations between the positive–negative, input–output, and sufficiency–necessity distinctions are such that (a) +Itests are always Stests, (b) −Itests are always Ntests, (c) +Otests are Stests for hypothesized positive relations and are Ntests for hypothesized negative relations, and (d) −Otests are Ntests for hypothesized positive relations and are Stests for hypothesized negative relations.

Given a primary +test bias, a secondary Stest bias can explain why $E(|\omega_B|) > E(|\omega_C|)$—namely, the remaining part of the CWI that is unexplained by a +test bias. First consider +tests, which should be more viable than −tests: If people display an Stest bias, then +Itests, which contrast A and B, should be more viable than +Otests, which contrast A and C, because the former are consistently associated with Stesting whereas the latter are not. Thus, under +test conditions, $E(|\omega_B|) > E(|\omega_C|)$. Now consider −tests: If people display an Stest bias, then −Itests, which contrast D and C, should be *less* viable than −Otests, which contrast D and B, because the former are consistently associated with Ntesting whereas the latter sometimes provide sufficiency information. Thus, under −test conditions, once again, $E(|\omega_B|) > E(|\omega_C|)$. Taken together, a primary +test bias and a secondary Stest bias should result in a positive–negative × input–output interaction: in terms of rule viability, +Itests > +Otests > −Otests > −Itests.

Experiment 1: Judgment of Causal Contingency

Several theories of causal inference posit that contingency is an important and perhaps necessary cue to causality (e.g., Cheng, 1997; Cheng & Novick, 1992; Kelley, 1967; Morris & Larrick, 1995; Schustack & Sternberg, 1981; Shanks & Dickinson, 1987; Wasserman, Elek, Chatlosh, & Baker, 1993). In Experiment 1, participants are required to integrate contingency information in the course of judging the magnitude and direction of the effect of an input on an output. We

---

accuracy is imperfect and intervening factors may influence the relation between two events. For instance, in everyday causal inference abnormal disabling conditions can intervene in what would otherwise have been a normal cause–effect relation (Cummins, Lubart, Alksnis, & Rist, 1991). Under such conditions, sufficiency and necessity are judged in relative terms (Einhorn & Hogarth, 1986; Jenkins & Ward, 1965; Suppes, 1970).

Hypothesized Positive Relation

	O	$\sim O$	
I	**Cell A** +,+ Confirmation	**Cell B** +,− Insufficiency	+Itest = Stest
$\sim I$	**Cell C** −,+ Nonnecessity	**Cell D** −,− Confirmation	−Itest = Ntest
	+Otest = Ntest	−Otest = Stest	

Hypothesized Negative Relation

	O	$\sim O$	
I	**Cell A** +,+ Insufficiency	**Cell B** +,− Confirmation	+Itest = Stest
$\sim I$	**Cell C** −,+ Confirmation	**Cell D** −,− Nonnecessity	−Itest = Ntest
	+Otest = Stest	−Otest = Ntest	

Figure 2. Contingency tables depicting the relations between positive–negative (+ vs. −), input–output (I vs. O), and sufficiency–necessity (S vs. N) distinctions by hypothesized relation direction. Otest = output test; Ntest = necessity test; Itest = input test; Stest = sufficiency test.

anticipate that, in this judgment domain, participants will display strong +test and Stest biases. Because there is great latitude in how the term *effect* may be defined, information-processing biases should have a strong influence on how people use contingency information to judge causality. In turn, we anticipate that lay conceptions of causality will reflect these biases.

Method

Participants read that the experiment was designed to study how people judge the relations between events, and that they would be asked to judge the effect of one event on another on the basis of the information they would receive. They were instructed to consider the contingency information carefully (but not to spend too much time trying to analyze it) and then to answer the question that followed. To minimize potential errors that are not directly related to statistical estimation (e.g., errors due to memory demands), we used a summary format for presenting contingency information. Each cell was described in a sentence (e.g., for Cell A the description was "the 'input' occurred and then the 'output' occurred") with the corresponding frequency (e.g., 20) printed next to it under a header labeled *number of instances out of "N"* (where *N* equals the sum of the four cell frequencies). Order of information presentation was counterbalanced so that half of the sample received information from Cells A, then B, then C, then D, and the other half received information in the reverse order.

Participants were given 36 information patterns (stimuli) to judge. After each stimulus, participants were asked, "How would you characterize the effect of 'the input' on 'the output'?" They responded on a 7-point scale ranging from -3 (*"the input" is a strong preventor of "the output"*) through 0 (*"the input" neither prevents nor causes "the output"*) to $+3$ (*"the input" is a strong cause of "the output"*). Absolute scale values of 1 and 2 were anchored at *weak* and *moderate*, respectively. Following the rating task, participants were asked if they had ever taken a statistics course. They also were asked whether they (a) consciously used a particular strategy, (b) consciously used different strategies depending on the information provided, or (c) often just guessed. Participants who indicated that they used a strategy were asked to describe it. Participants in the abstract conditions were asked if they mapped the abstract events onto more meaningful events. Following the experiment, participants were debriefed and thanked.

Stimuli. The 36 stimuli used in Experiment 1 are shown in Table 1. These stimuli consist of 12 frequency distributions, each of which is crossed with three different set sizes (i.e., the sum of the four frequencies either equaled 10, 20, or 40). For each quarter of the frequency distributions, one cell frequency is larger than the other three frequencies by factors of either 2.5 or 5.0. Across the stimuli, the four cells are matched in mean frequency. Therefore, any differential impact of the frequencies on causal ratings must be

Table 1
Stimuli and Mean Causal Judgments in Experiment 1

					Set size				
		10				20		40	
No.	A	B	C	D	M	No.	M	No.	M
1	5	2	2	1	.48	13	.51	25	.59
2	5	2	1	2	.52	14	.62	26	.73
3	5	1	2	2	.58	15	.62	27	.74
4	1	5	2	2	$-.44$	16	$-.62$	28	$-.68$
5	2	5	2	1	$-.44$	17	$-.47$	29	$-.66$
6	2	5	1	2	$-.31$	18	$-.38$	30	$-.50$
7	1	2	5	2	$-.31$	19	$-.35$	31	$-.34$
8	2	1	5	2	$-.11$	20	$-.15$	32	$-.14$
9	2	2	5	1	$-.19$	21	$-.09$	33	$-.23$
10	2	2	1	5	.12	22	.12	34	.13
11	2	1	2	5	.08	23	.18	35	.19
12	1	2	2	5	$-.04$	24	$-.06$	36	$-.09$

Note. Italicized values denote stimulus subsets within each set size whose largest frequency is *A*, *B*, *C*, or *D*. No. = stimulus number.

due to participants' differential weighting of this information. Two randomly selected orders of stimulus presentation were used.

Results and Discussion

Descriptives. Sixty-seven percent of the sample had never taken a statistics course; 67% reported that they consciously used a judgment strategy; and only 15% in the abstract condition claimed to have mapped the abstract events onto other events. Mean causal judgments (shown in Table 1) both across and within stimuli did not differ reliably as a function of any of these three variables, or of gender, smallest $p = .12.$[5] Nor did mean causal judgments differ as a function of stimulus set size, $p > .15$. Thus, subsequent analyses were collapsed across these groupings. The grand mean of participants' judgments is $-.01$ ($SD = 0.42$). This statistic indicates that participants, on average, were not biased toward perceiving facilitative relations more readily than inhibitory relations. Rather, the grand mean accurately reflects the fact that the four mean cell frequencies were equal. This finding is consistent with Wasserman and Shaklee (1984) who found no evidence that positive relations were perceived more readily than negative relations when contingency information was presented in summary format (however, these authors did find that positive relations were perceived more readily when contingency information was presented along a discrete time-line; see also Erlick & Mills, 1967).

The cell weight inequality. To assess the CWI, we correlated each of the four cell frequencies with each participant's judgments.[6] Next, we Fisher-transformed this 4×120 matrix of values from r_{ij} to z_{ij}, which is our measure of rule viability. The viabilities for B and C were multiplied by -1 so that they would be in the same direction as A and D. We then analyzed the viabilities for the four frequencies using a 3×4 (Event Type $\times$ Cell) mixed analysis of variance (ANOVA). There were reliable main effects of event type, $F(2, 117) = 3.66, p < .03, MSE = 0.05$, and cell, $F(3, 351) = 195.47, p < .001, MSE = 0.03$. The interaction

---

[5] The ratings in Experiments 1 and 2 were divided by 3 to provide a possible range of -1 to $+1$.

[6] In this research area, it has been assumed that participants' representations of frequencies reflect the objective values presented to them. In other domains, however, it has been found that people represent numeric magnitude as a nonlinear function of the arithmetic scale (Rule, 1969). This function has been approximated by a logarithmic transformation (Banks & Hill, 1974; Moyer & Landauer, 1967) or by a power transformation with an exponent of about .67 (Banks & Hill, 1974). If participants do transform frequency data as such, then the transformed frequencies should correlate more strongly than the actual frequencies with their ratings. We examined log and power (exponent of .67) transformations of the four cell frequencies and found that the actual frequencies correlated either the same or better than either transformation. Thus, we used the actual frequencies to calculate rule output. Note also that we used the cell frequencies, not the relative frequencies (i.e., cell frequency divided by total frequency), to calculate rule output.

Table 2
Mean Viabilities of Cell Frequencies by Event Type in Experiment 1

Event type	Cell frequency				
	A	B	C	D	M_{type}
Abstract	.45	.42	.16	.06	.27
Content–natural	.51	.41	.13	.09	.29
Content–social	.54	.41	−.03	−.02	.22
M_{cell}	.50	.41	.09	.04	.26

Note. Viabilities for B and C were multiplied by −1.

effect also was reliable, $F(6, 351) = 4.90, p < .001$. As can be seen in Table 2, which shows the cell and marginal means for this analysis, the main effect of event type and the interaction effect are primarily attributable to the near-zero weight attached to C and D in the content-social condition. The interaction effect is also attributable to the stronger weight associated with A in the content conditions, particularly the content-social condition. These findings indicate that people may be more inclined to adopt a +test strategy when processing meaningful events. Still, the magnitudes of these effects are very weak in contrast to the magnitude of the main effect of cell. As can be seen in the column marginals of Table 2, the CWI (as well as the expected valences of these viabilities) was replicated.[7] Clearly, however, the greatest weighting difference was between |A, B| and |C, D|.

To provide a sense of individual differences in cell weighting, we ranked the viabilities of the four cell frequencies for each participant from 1 to 4 (in some cases there were ties). Table 3 shows the percentage of cell-frequency viabilities at each rank. The modal ranking for A-viabilities was 1 followed by 2, for B-viabilities it was 2 followed by 1, and for C- and D-viabilities it was 4 followed by 3. Therefore, not only do people, on average, weight contingency information in a manner consistent with the CWI but most people weight this information in this manner. We also classified participants in terms of whether or not their judgments correlated reliably at $\alpha = .01$ with each of the four frequencies. The judgments of 80% of participants reliably correlated with A, and this figure declined to 63% for B, 10% for C, and 3% for D. Examining the results in

Table 3
Percentage of Cell-Frequency Viabilities by Magnitude Ranking in Experiment 1

Rank	Viability of cell frequency (%)			
	A	B	C	D
1	65	30	5	6
2	25	*54*	17	6
3	7	14	36	41
4	3	2	*42*	*47*

Note. Row averages do not equal 100% because of ties that increase the proportion of higher ranks. Italicized values identify the most frequent ranking for each cell-frequency viability.

Table 4
Mean Rule Viabilities by Positive–Negative (PN) and Input–Output (IO) Distinctions in Experiment 1

IO	PN		
	+test	−test	M_{IO}
Itest	.81	.11	.46
Otest	.47	.36	.42
M_{PN}	.64	.24	.44

more detail, the percentages of participants whose judgments were reliably correlated with various groupings of cell frequencies were as follows (in descending order): A only (37%), A and B only (35%), B only (12%), no frequency (6%), and each of the remaining categories accounted for less than 5%. Clearly, these idiographic analyses indicate that the vast majority of participants rely most heavily on Cell A and (to a lesser degree) on Cell B whereas, to most, Cells C and D are of relatively much less importance.

Positive-test and sufficiency-test biases. To examine our two main predictions, we calculated viabilities for four special cases of weighted ΔD: +Itest in which $\omega_A = \omega_B = 1$ and $\omega_C = \omega_D = 0$; +Otest in which $\omega_A = \omega_C = 1$ and $\omega_B = \omega_D = 0$; −Itest in which $\omega_C = \omega_D = 1$ and $\omega_A = \omega_B = 0$; and −Otest in which $\omega_B = \omega_D = 1$ and $\omega_A = \omega_C = 0$. These viabilities, the means of which are shown in Table 4, were then subjected to a series of planned comparisons. Supporting our prediction of +test bias, the +Itest was reliably more viable than the −Itest, paired $t(119) = 19.39, p < .001, SE = 0.04$. This comparison, however, is not a conservative test of the hypothesis because the +Itest also is expected to exceed the viability of the −Itest because of a possible Stest bias. That is, the former is consistently associated with Stesting whereas the latter is never associated with Stesting. A more conservative test is to compare the viability of the +Otest and the −Otest, both of which sometimes provide sufficiency information. As anticipated, the +Otest was reliably more viable than the −Otest, paired $t(119) = 3.79, p < .001, SE = 0.03$. The fact that the former contrast yielded a stronger effect suggests that people may jointly tend toward +testing and Stesting. To test for Stest bias in a manner that is unconfounded by the positive–negative distinction, we calculated two additional comparisons. As anticipated, the +Itest (which always provides sufficiency information) was reliably more viable than the +Otest (which sometimes provides sufficiency information), paired $t(119) = 11.84, p < .001, SE = 0.03$. Also as expected, the −Otest (which sometimes provides sufficiency information) was reliably more viable than the −Itest (which never provides sufficiency information), paired $t(119) = 11.62, p < .001, SE = 0.02$. Taken together, these findings support the notion that people's information integration strategies in causal inference are biased toward +testing and Stesting.

---

[7] For brevity, then, we collapsed over event type in subsequent analyses.

Our findings indicate that of the four linear contrasts constructed by crossing the positive–negative and input–output feature distinctions, the +Itest is the most viable. Although this follows from our predictions of +test and Stest biases, it may be that an even simpler strategy such as the Cell A strategy, or a strategy that is unclassified in terms of the positive–negative or input–output distinctions such as the confirming-cases strategy (viz., $A + D$), provides a better model. The +Itest ($\bar{z} = .81$) was reliably more viable than the Cell A strategy ($\bar{z} = .50$), paired $t(119) = 17.61$, $p < .001$, $SE = 0.02$. In turn, the Cell A strategy was reliably more viable than the confirming-cases strategy ($\bar{z} = .34$), paired $t(119) = 10.19$, $p < .001$, $SE = 0.02$. Using a nonmetric strategy-classification procedure, Shaklee and Wasserman (1986) found the same ordering for these three rules. The fact that +Itesting outperformed both the Cell A and confirming-cases strategies suggests that people assess trade offs in specific sources of information that are at odds in terms of their support for a target hypothesis rather than simply seeking confirmatory information.

Concepts of causation. We asked an additional 52 undergraduate participants from UBC to provide their definitions of the terms *cause* and *preventor*. If participants are biased toward Stesting, then they should be more likely to define *causes* and *preventors* in terms of sufficiency rather than necessity. For instance, participants should be more likely to state something like "when the cause is present the effect (likely) will occur" (i.e., the cause is sufficient to produce the effect) rather than something like "if the cause is not present the effect (likely) won't occur" (i.e., the cause is necessary to produce the effect). Given that people tend to think in terms that cohere with temporal order (Tversky & Kahneman, 1980), a sufficient cause should take the form $I \supset O$, whereas a necessary cause should take the form $\sim I \supset \sim O$. A sufficient preventor should take the form $I \supset \sim O$, whereas a necessary preventor should take the form $\sim I \supset O$.

First, we found that all participants stated their definitions in a manner that was consistent with temporal order. If statements were expressed in conditional form, the cause was always the implicans and the effect was always the implicate. Second, as we anticipated, 71% of participants defined *cause* strictly in $I \supset O$ (sufficiency) terms and 76% defined *preventor* strictly in $I \supset \sim O$ (sufficiency) terms. In contrast, only 22% and 10% of participants provided definitions of *cause* and *preventor*, respectively, that indicated that a necessity criterion was being considered. These findings cohere with the notion that people are biased toward Stesting. The findings also indicate that Stest bias is consistent with the primary meaning that people ascribe to causality and preventability.

Self-reports of strategy use. With the original sample of 120, we examined participants' self-reports of strategy use to see whether the weighting of cell information in their actual judgments conformed to the weighting specified in their reports. Of the 104 participants who claimed to have used a strategy, we isolated a subsample of 41 participants who claimed to use strategies that entailed an allocation of equal weight to all four cells when taken across the entire stimulus set.[8] Figure 3 shows the mean viabilities of the four cell
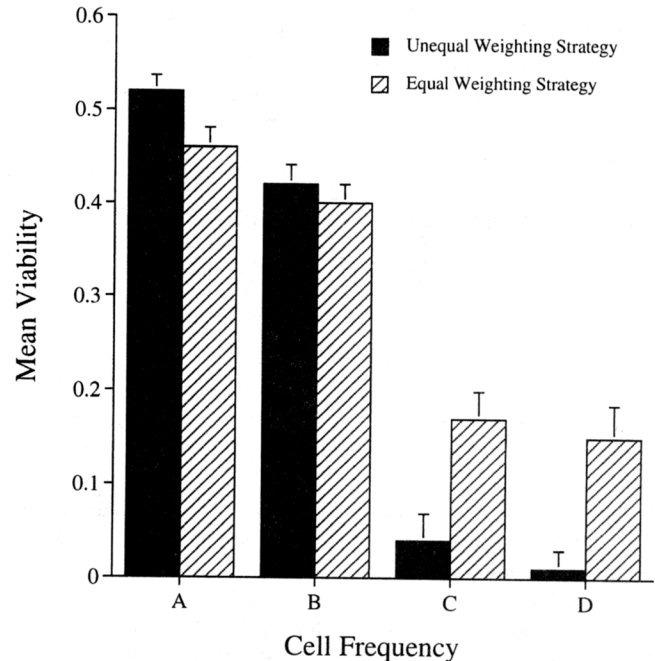

Figure 3. Mean viabilities of cell frequencies ($+SE$) for subsamples reporting to use equal or unequal cell-weighting strategies in Experiment 1. Viabilities for B and C were multiplied by -1.

frequencies between the *equal-weight* and *unequal-weight* subsamples of self-reported strategy users. The equal-weight subsample allocated weight more equally to the four cell frequencies than did the unequal-weight subsample. Specifically, they allocated reliably less weight to A, $t(102) = 2.01$, $p < .05$; reliably more weight to C, $t(102) = 2.82$, $p < .007$; and reliably more weight to D, $t(102) = 4.12$, $p < .001$ (there was no reliable difference between these groups for B, $t < 1$). Nevertheless, as can be seen in Figure 3, equal-weight participants, in fact, did not allocate equal weight to the four cell frequencies, $F(3, 120) = 38.95$, $p < .001$, $MSE = 0.03$. The unequal weighting was consistent with the CWI, attributable primarily to the greater weights associated with A and B than with C and D. These findings indicate that +test and Stest biases may persist even when people explicitly claim to use strategies that would eliminate these biases; however, under such circumstances, these biases are attenuated.

---

[8] Twenty participants described strategies that involved calculating either ΔP or $A/B - C/D$. Twelve participants described using either an anchoring or anchor-and-adjust strategy: The former strategy entailed considering only the cell with the largest frequency whereas the latter strategy entailed additional adjustments in view of the remaining cell frequencies. These strategies would lead to equal cell weighting because each cell was identical in terms of its distribution of frequency magnitudes across the entire stimulus set. Seven participants described using a linear combination of the four cell frequencies and the remaining two participants described using a linear combination of the four joint probabilities. Participants who mentioned using all four cells but who also explicitly mentioned an unequal allocation of cell weights were excluded from this subsample.

Experiment 2: Judgment of Noncausal Contingency

Most experimental contingency-judgment tasks include features that are suggestive of a causal relation between the target events. For instance, the events often are represented as antecedents and consequents, which itself may be taken as a cue to causality (Einhorn & Hogarth, 1986). Researchers often use events that world knowledge suggests are causally related (e.g., use of fertilizer and subsequent plant blooming; taking a drug and then developing a rash). And, the questions posed to participants often suggest a causal relation (e.g., when terms such as *effect* or *control* are used). In Experiment 2, we minimized the chances that participants would view the task as one involving causal judgment, allowing for a meaningful comparison with Experiment 1. We anticipated that, as in Experiment 1, participants would display +test and Stest biases owing to the fact that the term *relation* (like *effect*) is sufficiently general to allow information-processing biases to dominate.

Method

Forty undergraduates from UBC, different than those in Experiment 1, participated for course credit. Participants were given the same 36 stimuli used in the abstract condition of Experiment 1. Following each stimulus, participants were asked, "How would you characterize the relation between Event A and Event B?" They responded on a 7-point scale ranging from −3 (*strong negative relation*) through 0 (*no relation*) to +3 (*strong positive relation*). Contingency information was presented as in Experiment 1, except the word *then* was removed from each statement so as not to suggest a causal relation between the events. Stimuli were presented in the same two random orders as in Experiment 1. Once again, half the sample received information from Cells A, then B, then C, then D, and the other half received information in the reverse order. Following the rating task, participants were reminded that they received four types of frequency information (viz., the frequencies in Cells A to D) in the preceding task. They then were asked, "Which of these four types of information do you think are *necessary* to consider when assessing the relation between Event A and Event B?" Participants were told that they could circle as many types of information as they saw fit but that they must circle only those that they thought were necessary. Following the task, participants were debriefed and thanked.

Results and Discussion

The grand mean of participants' contingency judgments was −.02 (*SD* = 1.01). Thus, as in Experiment 1, participants, on average, did not exhibit a tendency toward perceiving positive relations more readily than negative relations. As well, as in Experiment 1, mean contingency judgments did not differ as a function of gender or stimulus set size.

The cell weight inequality. We analyzed the viabilities of each of the four cell frequencies as in Experiment 1. A one-way repeated-measures ANOVA revealed a reliable main effect of cell, $F(3, 117) = 45.01$, $p < .001$, $MSE = 0.02$. As in Experiment 1, the CWI was replicated: The mean viabilities for A to D were .41, .22, .16, and .01, respectively (once again, viabilities for B and C were multiplied by −1). Contrasting these findings with those from the abstract condition in Experiment 1, we see that participants attached

considerably less weight to B when judging noncausal contingency than when judging causal contingency ($\bar{z} = .22$ vs. .42, $\Delta = .20$). However, the absolute values of the difference in mean viability across the two judgment domains were negligible for A ($\Delta = .04$), C ($\Delta = .00$), and D ($\Delta = .05$). These findings indicate that given the same abstract events and the same frequency data, participants' judgments of causal contingency are more sensitive to statistical measures of contingency such as ΔP than are participants' judgments of noncausal contingency (but see Shaklee & Elek, 1988). Indeed, ΔP correlated .80 (.79 for phi) with the abstract group's mean ratings in Experiment 1 versus .71 (same for phi) with participants' mean ratings in Experiment 2.

As in Experiment 1, we ranked the viabilities of the four cell frequencies for each participant in order to provide a sense of individual differences in cell weighting. As Table 5 shows, the modal ranking for A-viabilities was 1, for B-viabilities it was 2, for C-viabilities it was 3, and for D-viabilities it was 4. We also found that the judgments of 45% of participants were reliably correlated ($\alpha = .01$) with A, and this figure decreased to 15% for B, and 8% for C and for D. As in Experiment 1, then, most participants tended to weight the cell frequencies in a manner consistent with the CWI.

Positive-test and sufficiency-test biases. As in Experiment 1, we examined the viabilities of four special cases of ΔD: +Itest in which $\omega_A = \omega_B = 1$ and $\omega_C = \omega_D = 0$; +Otest in which $\omega_A = \omega_C = 1$ and $\omega_B = \omega_D = 0$; −Itest in which $\omega_C = \omega_D = 1$ and $\omega_A = \omega_B = 0$; and −Otest in which $\omega_B = \omega_D = 1$ and $\omega_A = \omega_C = 0$. Table 6 shows the mean viabilities of these rules. Once again, supporting our prediction of +test bias, the +Itest was reliably more viable than the −Itest, paired $t(39) = 9.94$, $p < .001$, $SE = 0.04$. And, the +Otest was reliably more viable than the −Otest, paired $t(39) = 6.68$, $p < .001$, $SE = 0.04$. Once again, the fact that the former contrast yielded a stronger effect suggests that people may jointly tend toward +testing and Stesting. Supporting our prediction of Stest bias, the +Itest was reliably more viable than the +Otest, paired $t(39) = 2.43$, $p = .020$, $SE = 0.02$. And, the −Otest was reliably more viable than the −Itest, paired $t(39) = 2.49$, $p < .020$, $SE = 0.02$. As in Experiment 1, these findings support the notion that people's information integration strategies in causal inference are biased toward +testing and Stesting. Finally, as in Experiment 1, the +Itest ($\bar{z} = .52$) was reliably more viable than A

Table 5

Percentage of Cell-Frequency Viabilities by Magnitude Ranking in Experiment 2

	Viability of cell frequency (%)			
Rank	A	B	C	D
1	*83*	15	5	5
2	15	*43*	32	10
3	2	35	*48*	15
4	0	7	15	*70*

Note. Row averages do not equal 100% because of ties that increase the proportion of higher ranks. Italicized values identify the most frequent ranking for each cell-frequency viability.

Table 6
Mean Rule Viabilities by Positive–Negative (PN) and Input–Output (IO) Distinctions in Experiment 2

IO	PN		
	+test	−test	M_{IO}
Itest	.52	.14	.33
Otest	.47	.19	.34
M_{PN}	.50	.17	.34

alone ($\bar{z}$ = .41), paired $t(39)$ = 3.98, p < .001, SE = 0.03. In turn, A alone was reliably more viable than the confirming-cases strategy ($\bar{z}$ = .27), paired $t(39)$ = 4.79, p < .001, SE = 0.03. This finding adds further support to the notion that people assess tradeoffs between confirmatory and disconfirmatory sources of information rather than simply seeking confirmation.

Judgments of information necessity. Judgments of the types of information that are necessary for assessing a relation between two events were inconsistent with the actual weighting of these information types in the preceding judgment task. Table 7 summarizes participants' judgments of information necessity. As can be seen, the modal pattern was for participants to indicate that both Cells A and D were necessary to consider (22%). Examining the percentages of participants mentioning each of the cells, we found that the modal response was for participants to indicate that it is necessary to consider Cell A (68%). The next most frequently mentioned cell was D (50%), followed by B (48%) and then C (40%). Six two-tailed McNemar (binomial) tests were calculated to test the reliabilities of the pairwise differences in these percentages. The percentage mentioning Cell A was reliably greater than the percentage mentioning Cell C (p < .04) and marginally greater than the percentage mentioning Cell B (p < .08). The remaining differences in proportion were unreliable. Notably, although more than half of the sample indicated that Cell D was necessary to

Table 7
Percentage of Participants Indicating Various Patterns of Necessary Information in Experiment 2

Pattern	%
A	7.5
B	2.5
C	0.0
D	2.5
AB	13.0
AC	5.0
AD	22.5
BC	10.0
BD	2.5
CD	10.0
ABC	7.5
ABD	5.0
ACD	0.0
BCD	0.0
ABCD	7.5
∅	5.0

consider, D received the least amount of weight in the preceding judgment task. For 70% of the sample, D had the lowest viability of the four frequencies.

To further examine the correspondence between participants' reports of necessary information and their covariation judgments, we examined the mean viabilities of the four cell frequencies as a function of whether or not participants indicated that a particular frequency was necessary to consider (see Figure 4). Presumably, the viabilities should be of greater magnitude for the group indicating the necessity of using specific cell information (the necessity group) than for the group who did not indicate the necessity of using that information (the nonnecessity group). The mean viability of A was reliably greater for the necessity group than for the nonnecessity group, $t(36)$ = 3.68, p < .002. And the same was true for the mean viability of B, $t(36)$ = 2.35, p < .03. However, there was no reliable difference in mean viability between these groups for either C or D, ps > .14. Taken together, these findings indicate that what participants say is important to consider when judging contingency is not a good predictor of what they treat as important when actually doing so (see also, Shaklee & Hall, 1983). Indeed, the weight attached to A by participants who indicated that Cell A was not necessary to consider was greater than the weight attached to B, C, or D by participants who did indicate the necessity of the three corresponding cells. These findings add to the literature (e.g., Goldberg, 1968; Nisbett & Wilson, 1977) demonstrating that experts and nonexperts alike often are unable to describe their bases of judgment, many of which are likely to be automatic (Bargh, 1997).
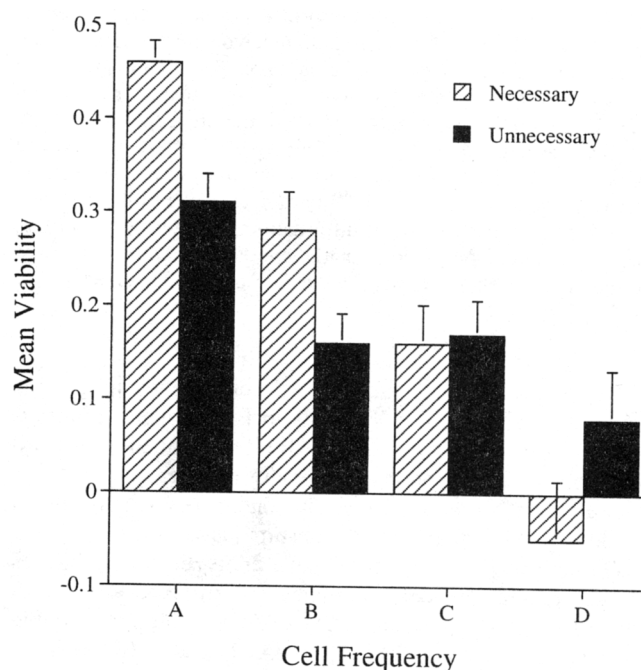

Figure 4. Mean viabilities of cell frequencies (+ SE) for groups indicating that a particular cell was either necessary or not necessary to consider in Experiment 2. Viabilities for B and C were multiplied by −1.

Experiment 3: Judgment of Conditional Probability

In Experiment 3, we examined whether the +test and Stest biases observed in causal and noncausal contingency judgment also would appear in conditional probability judgment. An important difference between contingency and probability judgment is that the latter (but not the former) has a single normative criterion that is evident. When someone is asked to judge how likely Y is to occur, given that X occurred, it is clear that the normative criterion is $P(Y|X)$. Most would agree that, when asked to judge causality, it is not similarly clear that the normative criterion is ΔP, or that the phi coefficient is the normative criterion for assessing noncausal contingency. Indeed, even among researchers, there is much debate concerning what normative causal judgment should mean (e.g., Anderson & Sheu, 1995). Given the transparency of norms for probability judgment, we expected that +test and Stest biases would be minimized when they were at odds with the normative criterion.

Method

One hundred and twenty-three undergraduates from UBC, different than those in Experiments 1 and 2, participated for course credit. Participants were randomly assigned to one of four conditions that differed only in terms of the conditional probability that was to be judged. Each probability corresponded to a particular positive–negative × input–output conjunction: In the +I condition, participants were asked to judge $P(O|I)$, which is the true-positive rate. In the −I condition, participants were asked to judge $P(O|{\sim}I)$, which is the false-negative rate. In the +O condition, participants were asked to judge $P(I|O)$, which is the hit rate. Finally, in the −O condition, participants were asked to judge $P(I|{\sim}O)$, which is the false-alarm rate.

Participants were presented with a modified version of the introduction to the content-social condition used in Experiment 1.[9] Participants were then instructed to assess, on the basis of the data presented, one of the four aforementioned probabilities. For instance, in the +I condition, they were asked, "How likely are people to start a conversation given that they had personality trait X?" Participants responded on an 11-point scale ranging from 0 (*not at all likely*) to 10 (*totally likely*). Contingency information was presented in a standard 2 × 2 table with a frequency value in each cell. Nine stimuli were used: For Cells A to D, respectively, the frequencies for these stimuli were (1) 6, 6, 6, 6, (2) 18, 2, 2, 2, (3) 2, 18, 2, 2, (4) 2, 2, 18, 2, (5) 2, 2, 2, 18, (6) 10, 10, 2, 2, (7) 2, 2, 10, 10, (8) 10, 2, 10, 2, and (9) 2, 10, 2, 10. Stimuli were presented in one of two randomly generated orders.

Results and Discussion

Normativeness of judgment. For each participant, the viability of each of the four probabilities was calculated. Table 8 shows the mean viabilities as a function of the positive–negative and input–output distinctions within each probability condition. Table 8 shows two important findings concerning the normativeness of judgment. First, in each condition, the normative probability (italicized value) was the most viable model of participants' judgments. That is, in each condition, the normative probability had a greater mean viability than the second most viable probability, which was

Table 8
Mean Rule Viabilities by Positive–Negative (PN) and Input–Output (IO) Distinctions for Each Probability Condition in Experiment 3

IO	PN +test	PN −test	M_{IO}
+Itest condition (n = 33)			
Itest	*1.23*	−.09	.57
Otest	.61	.07	.34
M_{PN}	.92	−.01	.46
−Itest condition (n = 32)			
Itest	−.28	*1.78*	.75
Otest	.49	.04	.27
M_{PN}	.11	.91	.51
+Otest condition (n = 30)			
Itest	.68	−.01	.34
Otest	*.85*	.10	.48
M_{PN}	.77	.05	.41
−Otest condition (n = 28)			
Itest	.51	.11	.31
Otest	−.13	*.79*	.33
M_{PN}	.19	.45	.32

Note. The normative probability within each condition is italicized.

not normative. These differences in mean viability were reliable in the +I condition, paired $t(32) = 2.76, p < .010$, $SE = 0.22$, and in the −I condition, paired $t(31) = 4.44, p < .001$, $SE = 0.29$, but were unreliable in the remaining two conditions, $ps > .30$. Second, each of the four probabilities was most viable in the condition in which it was normative. The difference between the mean viability of a probability in the condition in which it was normative and the next highest mean viability for that same probability was reliable in the +I condition, $t(61) = 2.34, p < .022, SE = 0.23$; the −I condition, $t(58) = 5.27, p < .001, SE = 0.32$; and the −O condition, $t(56) = 3.55, p = .001, SE = 0.25$; but was unreliable in the +O condition, $t(61) = 1.42, p = .159$, $SE = 0.17$. In general, then, participants' statistical estimates of likelihood conformed to the normative criterion of what they were asked to judge.

We also examined individual differences in the degree to which participants' judgments reflected the normative criterion in their condition. Table 9 shows, as a function of condition, the percentage of participants whose judgments correlated with the normative probability in their condition within various magnitudinal ranges. Two descriptive findings are noteworthy: First, over half of participants in the −I condition exhibited correlations above .90. Second, roughly 20% of participants in the −O condition provided judgments that correlated negatively with the normative probability. To

---

[9] The last two lines were replaced by the sentence "You then summarized the results individually for each of the personality traits that you were interested in examining."

Table 9

Percentage of Participants Showing Particular Correlations Between the Normative Probability and Their Judgments by Condition in Experiment 3

r interval	Condition (%)			
	+I	−I	+O	−O
<0	3	3	3	*21*
0–.1	6	3	10	0
.1–.2	9	3	3	7
.2–.3	0	3	0	11
.3–.4	3	0	13	11
.4–.5	0	6	3	18
.5–.6	3	3	13	0
.6–.7	21	3	20	4
.7–.8	27	16	7	7
.8–.9	6	6	17	4
.9–1.0	21	*53*	10	18

Note. Italicized values indicate the interval capturing the largest percentage of participants per condition. I = input; O = output.

compare the degree to which the preceding nomothetic account provides a reasonable summary of individual participants' judgments as a function of probability condition, we contrasted the highest percentages in each condition for three interval sizes taken from Table 9. For instance, using a single interval of r to r + .10, 27%, 53%, 20%, and 21% of participants were classified in +I, −I, +O, and −O conditions, respectively. Similar comparisons were made for intervals two and three times greater than the preceding interval. The reliabilities of the differences between the percentages obtained using these three classifications are listed in Table 10. For each classification, the greatest percentage of participants was clustered together in the −I condition, indicating that the previous nomothetic account best represents participants asked to judge $P(O|{\sim}I)$.

Positive-test and sufficiency-test biases. We analyzed the viabilities of the normative probability for each condition using a set of planned comparisons comparable to that used in the preceding experiments. Unlike the findings of Experiments 1 and 2, normative +tests did not differ reliably from normative −tests in terms of their viability; this was true for both the comparison of the +Itest and the −Itest, $t(53) = 1.66, p = .102, SE = 0.33$, and the comparison of the +Otest and the −Otest, $t < 1$. Nor did an analysis of the viabilities of the normative probabilities yield support for an Stest bias: The +Itest did not differ reliably from the +Otest, $t(61) = 1.47, p = .146, SE = 0.26$, and the −Itest was reliably *more* viable than the −Otest, $t(58) = 2.79, p = .007, SE = 0.36$. These findings support the notion that +test and Stest biases can be attenuated or even eliminated when judgments are guided by a transparent normative criterion that is at odds with +testing or Stesting.

Participants' probability judgments were no more accurate for normative +tests than for normative −tests, as one might expect given the clarity of the normative criteria in this task. Nevertheless, there is strong evidence for a residual +test bias. In each condition, regardless of whether the normative probability was a +test or a −test, the most

viable nonnormative probability was always a +test. Namely, the most viable nonnormative probabilities were the +Otest in the +I and −I conditions and the +Itest in the +O and −O conditions. In each condition, these nonnormative +tests are reliably more viable than the next most viable nonnormative probability, all $ps < .001$ for paired t tests. These findings suggest that the impact of +test bias is suppressed for judgments guided by transparent normative criteria at odds with +testing, despite the fact that the bias continues to residually exert an impact on information processing. The suppression rather than the elimination of +test bias provides a better characterization of the overall set of findings from this experiment.

General Discussion

A robust finding in the contingency-judgment literature, termed the CWI, is that people assign the greatest importance to Cell A followed in order by Cells B, C, and D. We accounted for the CWI in terms of a primary +test bias and a secondary Stest bias. We demonstrated the predictive superiority of +tests over −tests and of Stests over Ntests in the domains of causal and noncausal contingency judgment. The finding that the vast majority of participants defined the terms *cause* and *preventor* exclusively in terms of a sufficiency criterion also supports the notion that people display an Stest bias. However, we also demonstrated that, when a transparent normative criterion is at odds with +testing or Stesting, the impact of +test and Stest biases is minimized or suppressed. This result, found in Experiment 3, is that judgments may tend toward normativeness despite the background effect of these information-processing biases.

Biases of Computation Versus Criterion Selection

In a typical contingency-judgment experiment, participants are not asked to judge specific probabilities but rather to assess the effect of one event on another or to assess the relation between these events. Terms such as *control, effect,* or *relation,* however, leave considerable room for interpretation by participants. Therefore, participants not only have to compute contingency, they have to select subjectively appropriate criteria for assessing contingency, given their understanding of the construct at hand. This suggests two possible types of bias: *criterion-selection bias,* which entails

Table 10

Percentages of Participants Classified for Three Correlation Intervals in Experiment 3

Interval criterion	Condition (%)			
	+I	−I	+O	−O
r to r + .10	27_a	53_b	20_a	21_a
r to r + .20	$48_{a,b}$	59_a	33_b	29_b
r to r + .30	$54_{a,b}$	75_a	42_b	40_b

Note. Within rows, percentages not sharing a common subscript differ reliably at $\alpha = .05$ by pairwise chi-square tests $(df = 1, \chi^2_{critical} = 3.84)$. I = input; O = output.

an unequal weighting of potentially relevant judgment criteria, such as viewing Stesting as more important than Ntesting in judging contingency; and *computational error*, which entails weighting information in a manner that is suboptimal for testing one's subjective criteria, such as overweighting Stesting relative to Ntesting if one wishes to weight them equally.

Our findings indicate that the overweighting of +tests and Stests relative to their counterparts represents both criterion-selection bias and computational error. Because participants' computations in Experiment 1 generally cohered with lay conceptions of causality (which we assessed using an independent sample), it appears that one source of the biases reported stems from criterion selection. Findings from past research that has asked participants to state what types of contingency information are necessary for judging covariational relations also support this notion. For example, Crocker (1982) and Wasserman et al. (1990) found that the greatest proportion of their participants deemed Cell A as necessary, followed in order by Cells B, C, and D. However, in our Experiment 2, over half the sample indicated the necessity of considering Cell D despite the fact that those who did so ascribed relatively the least weight to this cell. Nor did these participants ascribe more weight to *D* than participants who did not indicate the necessity of considering Cell D. And, in Experiment 1, self-reported equal-weight strategy users still exhibited evidence of weighting frequency information in a manner consistent with the CWI. These findings indicate computational error because participants' weighting of information in judging contingency is incongruent with the types of information that they state are important for such judgments.

Prescriptive Issues

In everyday life, judgment strategies often reflect a tension between costs and benefits that are not always apparent in laboratory settings (Arkes, 1991; Friedrich, 1993; Funder, 1987). It is important to question whether it is appropriate to view +test and Stest biases as mistakes. Certainly, these biases exact costs under particular circumstances (e.g., in cases in which detecting false-negative errors is important). The question, however, is whether they are overly costly given the range of conditions to which people are exposed. Klayman and Ha's (1987) probability analysis of the diagnostic value of information indicates that +testing is an adaptive strategy. As noted earlier, Stest bias may also be adaptive because the failure to detect insufficiency often exacts greater practical costs than the failure to detect nonnecessity. If practical costs generally favor some types of error minimization over others, then differential cell weighting follows as a rational response. Moreover, the findings from Experiment 3, indicating that the impact of these biases is minimized when the task involves a transparent normative criterion, suggest that people are able to shift their focus when they perceive the need to do so (cf. Friedrich, 1993). Given a pragmatic rationale for favoring Stesting over Ntesting, ΔP and phi may be unfavorable as prescriptive models of causal and contingency judgments,

respectively. Indeed, any model that does not take both motivational and discriminability issues into account is likely to be prescriptively shortsighted (Killeen, 1981) and descriptively invalid (e.g., see Harkness, DeBono, & Borgida, 1985, who found that personal involvement influences how the four cells are weighted).

Normative Issues

The delta and phi rules typically are viewed as normative models of causal and contingency judgments, respectively. Given that the CWI clearly is inconsistent with the equal cell weighting of these rules, it is instructive to consider the meaning of normativeness. In our view, a normative principle must convey a mathematical or logical truth or there must be a clear mapping between verbal and mathematical definitions, as was the case in Experiment 3. An example of a normative principle is the conjunction rule, which asserts that $P(X \cup Y) \leq P(X) \cap P(Y)$; namely, within any target set, the probability of joint occurrence is less than or equal to the probabilities of the unconjoined elements. Whereas people may erroneously violate this rule in judging probability (Kahneman & Tversky, 1972), probabilities themselves are never in violation of the rule. The fact that probabilities obey the conjunction rule is what makes that rule normative. Nor should it be assumed that valid normative principles make good prescriptive principles given the total set of conditions describing the systems within which humans must function. Normative principles do not refer to how decision makers should ideally perform but to how they would perform if they could exist within the ideal realm of mathematics—a distinction that often is muddled in decision science.

In contrast to the conjunction rule, ΔP and phi do not specify mathematical truths. Rather, they are models that reflect specific processing objectives. It is true that these rules are established mathematical procedures for assessing one- and two-way dependency between binary variables, respectively. These conventions, however, do not serve as a valid basis for ascribing normative status to the rules. Even if the prescriptive value of these rules were to be demonstrated, it still would not make the normative argument defensible. We, therefore, suggest that it may be inappropriate to treat statistical measures such as ΔP and phi as absolute criteria for assessing the accuracy of causal and contingency judgments (contrast Shaklee, 1983).

Additional Descriptive Issues

Computational units. Although we have used a feature-analytic approach to examine why people unequally weight the four types of contingency information, we also can use this approach to examine the types of computational units that people tend to combine. We found that in both causal and noncausal contingency judgment, a +Itest strategy provided the best description of participants' judgments. But how is a +Itest computationally instantiated? We examined the viabilities of four possible instantiations: a frequency rule, $A - B$; a ratio rule, A/B; a conditional-probability rule, $A/(A + B)$; and a joint-probability rule, $(A - B)/N$. This set

of rules holds the type of contingency information (viz., Cells A and B) constant, while systematically varying the type of units in which it is expressed.

We found that, for causal judgment (Experiment 1), there was a reliable difference in the mean viabilities of these rules, $F(3, 351) = 71.81$, $p < .001$, $MSE = 0.01$. This difference is primarily attributable to the greater viability of the joint-probability rule ($\bar{z} = .90$) in contrast to the frequency ($\bar{z} = .81$), conditional-probability ($\bar{z} = .80$), and ratio ($\bar{z} = .73$) rules. Similarly, for noncausal contingency judgment (Experiment 2), simple contrasts revealed that the mean viability of the joint-probability rule ($\bar{z} = .53$) was reliably greater ($p = .001$) than that of the ratio rule ($\bar{z} = .49$) but did not differ reliably from either the frequency rule ($\bar{z} = .52$) or the conditional-probability rule ($\bar{z} = .50$). The better performance of joint-probability rules in Experiments 1 and 2 may be attributable to the fact that participants were provided with cell frequencies under a header labeled *number of cases out of "N."* This format likely heightened the salience of joint probabilities (see Anderson & Sheu, 1995, who argued that causal ratings are judgments of magnitude that reflect perceptually salient variables in a given context).

Recall that in Experiment 3 there were four conditions and that for each there was a normative contrast. For example, in the $+I$ condition, a contrast of A and B is instantiated as $P(O|I)$. We created three variables by calculating the viabilities of the appropriate contrast in each condition in three computational forms: frequency contrasts (which, in this experiment, are equivalent to joint-probability contrasts in predictive terms), ratios, and conditional probabilities. There was a reliable difference in the mean viabilities of these rules, $F(2, 244) = 27.13$, $p < .001$, $MSE = 0.22$. Simple contrasts revealed that the mean viability of the normative conditional probabilities ($\bar{z} = 1.18$) was reliably greater than that of both the frequency ($\bar{z} = 1.00$) and ratio ($\bar{z} = .74$) rules, $ps < .001$. Therefore, not only did participants tend to select the normatively appropriate contrast of cells, they also tended to instantiate it in the appropriate computational form given the task.

Output density bias. Several studies (see Allan, 1993, for a review) report that participants' contingency judgments are correlated with the frequency of output presence, $f(O) = A + C$—a finding termed *output density bias.* We have proposed that output density bias is primarily an artifact of $+$test bias and the valence of the cell weights. Given that ω_A is positive and ω_C is negative, the viability of $f(O)$ should agree with the following two inequalities in viabilities:

$$z_A \geq z_{A+C}. \tag{3}$$

$$z_{A-C} \geq z_{A+C}. \tag{4}$$

In support of Inequality 3, we found that $z_A = .50 > z_{A+C} = .25$ in Experiment 1 and that $z_A = .41 > z_{A+C} = .15$ in Experiment 2. Namely, the Cell A strategy should be more predictive of participants' judgments than output density, which inappropriately weights ω_C positively. In support of Inequality 4, we found that $z_{A-C} = .47 > z_{A+C} = .25$ in

Experiment 1 and that $z_{A-C} = .47 > z_{A+C} = .15$ in Experiment 2. Namely, a $+$Otest strategy is more predictive of participants' judgments than is output density. Moreover, given the CWI, which indicates that $|\omega_B| > |\omega_C|$, and the finding that both these weights are negative, the viability of output density should be greater than the viability of input density ($A + B$). This predicted inequality was found in Experiment 1 ($z_{A+C} = .25 > z_{A+B} = .05$) and to a lesser extent in Experiment 2 ($z_{A+C} = .15 > z_{A+B} = .11$). Taken together, these findings suggest that the predictive values of output density and other strategies can themselves be predicted on the basis of knowing how people weight information from the four cells. There is, however, nothing particularly notable about output density bias per se.

Future Research

In this article, we have introduced both a conceptual approach and a theoretical account of the CWI. These contributions hopefully will generate useful extensions to what we currently know about contingency-based judgment processes such as hypothesis testing, causal inference, and probability judgment to name a few. One direction for future research is to extend the current analysis of single-cue tasks to multiple-cue tasks (e.g., see Melz, Cheng, Holyoak, & Waldmann, 1993; Shanks, 1991). The latter often is the more ecologically valid case and, therefore, warrants attention. Another line of investigation might focus on the links between error-reduction goals, criteria selection, and the emergence or suppression of various information-processing biases (see Friedrich, 1993). For example, would providing affectively positive incentives (e.g., a higher rate of remuneration) or negative incentives (e.g., a threat of losing out on a desired outcome) for minimizing false-negative errors serve to attenuate $+$test and Stest biases? Such research is needed to develop an account that takes both discriminability and motivational issues into consideration.

In our research, we manipulated task objectives by examining contingency information integration in three different judgment domains. Within-domain manipulations of task objectives, however, also are needed. For instance, participants could be asked specific questions about either necessary or sufficient causation. This would permit additional tests of the notion that people display an Stest bias and the notion that the transparency of normative criteria influences the degree to which the impact of pre-existing biases may be suppressed. Additional tests of the notion that people display $+$test bias could be performed by manipulating whether the input and output variables are asymmetric or symmetric. For example, if the input variable is symmetric (e.g., Fertilizer 1 vs. Fertilizer 2) and the output variable is asymmetric (e.g., plants bloom or do not bloom), then, in terms of the positive–negative distinction, we can define the contrast of A and C as positive, the contrasts of A and B and of D and C as neutral, and the contrast of D and B as negative. Accordingly, we would expect that the weight given to A and C be greater than that given to B and D. Therefore, one also can make predictions on the basis of our account about when the CWI would not be found. The

manipulation of how variables are represented is also important in order to better understand possible differences in information integration across judgment domains. For instance, the asymmetric case may be more representative of situations in which people make causal- rather than non-causal-contingency judgments, and the reverse may be true for the symmetric case.

Concluding Remarks

A considerable amount of theorizing has taken place concerning the types of biases people display in top-down or theory-driven judgments of causation and covariation (e.g., Nisbett & Ross, 1980). By contrast, our understanding of people's bottom-up or data-driven strategies for integrating contingency information in the course of judgment is currently at a more descriptive stage. A specific goal here was to develop a theoretical account of one of the most robust findings in this area of research: the CWI. We provided a parsimonious account of the CWI in terms of two information-processing biases that have been demonstrated to influence responses to other types of cognitive judgments (e.g., concept attainment). A more general goal of this article was to offer a conceptual analysis of, and data-analytic response to, past rule-based approaches to characterizing contingency judgment, which have been largely descriptive and inconclusive. Our feature-analytic approach does not focus on finding the best predictor of participants' ratings. Rather, it focuses on making inferences about features of human information processing so that what Stone and Van Orden (1994) have termed *design principles*—principles that relate model behavior to observable human behavior—may be discovered.

References

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

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

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

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

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

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

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110,* 486–498.

Banks, W. P., & Hill, D. K. (1974). The apparent magnitude of number scaled by random production. *Journal of Experimental Psychology Monograph, 102*(2), 353–376.

Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer, Jr. (Ed.), *Advances in social cognition* (Vol. 10, pp. 1–61). Mahwah, NJ: Erlbaum.

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory & Cognition, 10,* 511–519.

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

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

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

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

Clark, H. H., & Clark, E. V. (1977). *Psychology and language.* New York: Harcourt Brace Jovanovich.

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

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

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19,* 274–282.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3–19.

Erlick, D. E., & Mills, R. G. (1967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology, 43,* 9–14.

Fazio, R. H., Sherman, S. J., & Herr, P. M. (1982). The feature-positive effect in the self-perception process: Does not doing matter as much as doing? *Journal of Personality and Social Psychology, 42,* 404–411.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3–32.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill.

Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review, 100,* 298–319.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75–90.

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

Halberstadt, N., & Kareev, Y. (1995). Transitions between modes of inquiry in a rule discovery task. *Quarterly Journal of Experimental Psychology, 48A,* 280–295.

Harkness, A. R., DeBono, K. G., & Borgida, E. (1985). Personal involvement and strategies for making contingency judgments: A stake in the dating game makes a difference. *Journal of Personality and Social Psychology, 49,* 22–32.

Hovland, C. L. (1952). A "communication analysis" of concept learning. *Psychological Review, 59,* 461–472.

Hovland, C. L., & Weiss, W. (1953). Transmission of information concerning concepts through positive and negative instances. *Journal of Experimental Psychology, 45,* 175–182.

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

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

Jennings, D. L., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211–230). New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. *Cognitive Psychology, 3,* 237–251.

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

Kareev, Y. (1995). Positive bias in the perception of covariation. *Psychological Review, 102,* 490–502.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15, pp. 192–238). Lincoln: University of Nebraska Press.

Killeen, P. R. (1981). Learning as causal inference. In M. L. Commons & J. A. Nevin (Eds.), *Quantitative analyses of behavior: Vol. 1. Discriminative properties of reinforcement schedules* (pp. 89–112). Cambridge, MA: Ballinger.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 596–604.

Levin, I. P., Wasserman, E. A., & Kao, S.-F. (1993). Multiple methods for examining biased information use in contingency judgments. *Organizational Behavior and Human Decision Processes, 55,* 228–250.

Lipe, M. G. (1990). A lens model analysis of covariation research. *Journal of Behavioral Decision Making, 3,* 47–59.

Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology, 71,* 450–463.

McGill, A. L., & Klein, J. G. (1993). Contrastive and counterfactual thinking in causal judgment. *Journal of Personality and Social Psychology, 64,* 897–905.

McGill, A. L., & Klein, J. G. (1995). Counterfactual and contrastive reasoning in explanations for performance: Implications for gender bias. In N. J. Roese & J. M. Olson (Eds.), *What might have been: The social psychology of counterfactual thinking* (pp. 333–352). Hillsdale, NJ: Erlbaum.

McGuire, W. J., & McGuire, C. V. (1992). Cognitive-versus-affective positivity asymmetries in thought systems. *European Journal of Social Psychology, 22,* 571–591.

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla–Wagner learning rule? Comment on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1398–1410.

Morris, W. M., & Larrick, R. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102,* 331–355.

Moyer, R. S., & Landauer, T. K. (1967). The time required for judgments of numerical inequality. *Nature, 215,* 1519–1520.

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

Nisbett, R. E., & 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 of mental processes. *Psychological Review, 84,* 231–259.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29–46.

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

Rule, S. J. (1969). Equal discriminability scale of number. *Journal of Experimental Psychology, 79,* 35–38.

Salmon, W. C. (1965). The status of prior probabilities in statistical explanation. *Philosophy of Science, 32,* 137–146.

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

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

Shaklee, H., & Elek, S. (1988). Cause and covariate: Development of two related concepts. *Cognitive Development, 3,* 1–13.

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

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 2,* 208–224.

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

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

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

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

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

Stone, G. O., & Van Orden, G. C. (1994). Building a resonance framework for word recognition using design and system principles. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1248–1268.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam: North-Holland.

Toffler, A. (1990). *Power shift: Knowledge, wealth, and violence at the edge of the 21st century.* New York: Bantam Books.

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

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (Vol. 1, pp. 49–72). Hillsdale, NJ: Erlbaum.

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

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology, 11,* 92–107.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 32,* 129–140.

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., 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., & Shaklee, H. (1984). Judging response–outcome relations: The role of response–outcome contingency,

outcome probability, and method of information presentation. *Memory & Cognition, 12,* 270–286.

Yates, J. F., & Curley, S. P. (1986). Contingency judgment: Primacy effects and attention decrement. *Acta Psychologica, 62,* 293–302.

Appendix

Introductory Text Used in the Content Conditions in Experiment 1

Introduction in the Content-Natural Condition

Suppose that you are employed by National Flowering Plants Laboratory (NFPL). NFPL has developed 36 experimental fertilizers, which are labeled F1 to F36, for promoting the Lanyu to bloom. The Lanyu, an exotic plant, is imported from Brazil. You will test these 36 fertilizers in a random order. For each fertilizer, you will study a completely different group of plants. Within each group, you will give the fertilizer to some plants but not to others. Aside from whether or not plants received fertilizer and the type of fertilizer they may have received, all plants were exposed to the same environmental conditions (for example, watering schedules, amount of sunlight). Because of the changing availability of the Lanyu, some groups consisted of 10 plants, others consisted of 20 plants, and some consisted of 40 plants. One month after testing, for each of the 36 fertilizer groups (Groups F1 to F36), you collect information about the number of cases for each of the following possibilities: (The four conjunctions are then typed out.)

Introduction in the Content-Social Condition

Suppose that you are interested in examining the effect of 36 different personality traits (which are labeled T1 to T36) on

people's willingness to start a conversation with a stranger. You devise a study to address this question. Participants in your study completed a personality test and were classified either as having a given personality trait or not having that trait. A participant would then be placed in a social situation involving a stranger (actually someone who was working for you). The stranger would never start a conversation with a participant and would just keep to herself. Participants were led to believe that the stranger (like themselves) was waiting for the next phase of the study to begin. In fact, you were simply recording whether a participant started a conversation with the stranger within a 5-min period. A separate group of participants was used to test each trait; these groups consisted of either 10, 20, or 40 participants depending on participant availability. You then summarized the results individually for each of the 36 traits as follows: (The four conjunctions are then typed out.)