FEATURES OF INFORMATION INTEGRATION IN CAUSAL INFERENCE

by

DAVID R. MANDEL

B.A., Concordia University, 1989
M.A., University of British Columbia, 1992

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Psychology

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September, 1996

Department of Psychology

The University of British Columbia
Vancouver, Canada

Date September 19, 1996

ABSTRACT

An important aspect of causal inference is assessing the contingency between antecedents and outcomes. Research on how people integrate contingency information has focused on identifying the "best" rule to descriptively model the information integration process. In contrast to this rule-analytic approach, the present feature-analytic approach asks the question, "What features are important in describing the information integration process?" Five key propositions of the present account are that (a) people prefer strategies that involve contrasting data with conflicting implications to strategies that involve seeking only confirmatory or marginal-frequency data, (b) people weigh positive information more heavily than negative information, (c) people are biased toward testing sufficiency rather than necessity, (d) people are biased toward strategies that cohere with the perceived direction of time (input tests) rather than those that violate this perception (outcome tests), and (e) people are biased toward probability strategies that enable comparability across data contexts rather than frequency strategies that do not. In three experiments, subjects received contingency information on two, temporally sequenced, binary variables in numeric summary format. Subjects were asked to rate the direction and magnitude of the causal relation between the two variables based on the contingency information provided. Results of Experiments 1-3, corroborated by a reanalysis of data from two published experiments employing a discrete-trial method for presenting stimuli, strongly supported the first four propositions. To test the fifth proposition, I reanalyzed data from five published experiments in addition to an analysis of data from Experiment 3. Results indicated that within each data context preferences for either frequency, conditional-probability, or joint-probability strategies emerged, but across contexts consistent preferences for one type of combination method was lacking. Taken together, the findings indicate that invariant properties of the information integration process in causal inference can be isolated but these consist of systematic feature preferences rather than stable rankings of rules in terms of their predictive utility.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

This research was supported in part by Doctoral Fellowships from the Natural Sciences and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada. In bringing this thesis to fruition several people also deserve thanks. I am indebted to Darrin Lehman, whose critical feedback, keen insight, generosity, and concern have made an invaluable contribution to my intellectual life at UBC. I also thank my thesis committee members, Eric Eich and Lawrence Ward, for providing important feedback that improved the quality of this thesis, and I thank other members of my examination committee, Peter Graf, Geoff Hall, Ziva Kunda, Charles Weinberg, and Donald Wehrung for taking the time to carefully read and comment on my thesis. Finally, I thank Sarah Smith for her deep friendship, which has helped to sustain me over the past 5 years.

To my wonderful parents,

Veronika Schwartz and Miklós Mandel

INTRODUCTION

In physical, social, and scientific realms, causal inferences allow people to explain the past, exert control in the present, and predict the future. In fact, many of our everyday activities depend on implicit causal knowledge derived through causal inference (Read, 1987), and language itself is saturated with implicit causal meaning (Brown & Fish, 1983; Garvey & Caramazza, 1974; Kasof & Lee, 1993). People are especially likely to search for causes when faced with negative or unexpected outcomes (Kanazawa, 1992; Pyszczynski & Greenberg, 1981; Taylor, 1991; Weiner, 1985). However, even in reasoning about noncausal relations (e.g., conditional or biconditional implications), people may rely on pragmatic rules of causal inference (Cheng, Holyoak, Nisbett, & Oliver, 1986), such as the causal schemata proposed by Kelley (1971, 1973). In short, causal inference allows people to bind together events in meaningful ways that answer self-posed questions of *how* and *why* things occur as they do (Hilton, 1990; Wong & Weiner, 1981). As philosopher David Hume (1740/1938) put it, causal thinking is the "cement of the universe."

Several factors have been proposed to influence causal inference; for example, temporal order, spatial and temporal contiguity, and similarity of cause and effect (for a review of these factors, see Einhorn & Hogarth, 1986), event abnormality (Hilton & Slugoski, 1986; Kahneman & Miller, 1986), and counterfactual availability (Lipe, 1991; Wells & Gavanski, 1989). One factor thought to be a necessary but insufficient component of causal inference is estimating the degree and direction of contingency between antecedent and consequent events (Anderson & Sheu, 1995; Cheng, 1993; Cheng & Novick, 1991, 1992; Schustack & Sternberg, 1981; Shaklee & Elek, 1988; Wasserman, Chatlosh, & Neunaber, 1983; Wasserman, Elek, Chatlosh, & Baker, 1993). As others (e.g., Alloy & Tabachnik, 1984; Crocker, 1981; Fiske & Taylor, 1991; Jennings, Amabile, & Ross, 1982) have noted, assessing contingency involves many stages including (a) deciding what kinds of data are relevant, (b) sampling cases, (c) classifying instances, (d) recalling evidence and estimating the frequencies of confirmatory and disconfirmatory instances, and finally (e) integrating (i.e., weighing and combining) the evidence to make a contingency judgment. It is the last of these stages—people's information integration strategies in the causal inference domain—that is the focus of the present research.

The literature on this topic has been shaped by two key objectives: (a) to identify the rule that provides the best summary description of people's information integration strategies and (b) to document departures of people's strategies from normative prescriptions. Here, *strategies* refer to the ways in which people go about weighing and combining contingency information. *Rules* refer to the specific formal models that researchers have focused on in their attempts to provide a descriptive summary of people's strategies. The aforementioned objectives rest on a set of implicit assumptions. For example, the search for a single rule that describes the integration process presupposes that a "best rule" might be found. This objective also presupposes that rule discrimination is an important focus of analysis. A primary goal of the present research is to consider whether these assumptions are tenable. In fact, I will argue that they are not.

In response to these issues, I develop an alternative framework for research on information integration that represents a fundamental shift away from the current approach. In contrast to what Allan (1993) termed the field's "quest for the best rule," a key aspect of the present framework is its emphasis on identifying important *features* of the information integration process. This emphasis is exemplified at many levels: At a paradigmatic level, I argue that a feature-analytic approach has distinct advantages over the traditional rule-analytic approach. At a methodological level, I develop the technique of *feature analysis*, which extends the correlational method as it is typically applied in this domain. Finally, at a substantive level, I argue that information integration in causal inference is characterized foremost by a contrastive, positive-test strategy in which people weigh data concerning the presence of an event more heavily than data concerning the absence of an event. Additionally, I suggest that people are biased toward sufficiency testing and toward test strategies that cohere with people's experience of time flowing forwards from antecedent to consequent events. In terms of *how* people integrate information, I consider the issue of integrating frequencies versus probabilities. The distinction is broadened to include joint-probability rules and it is discussed in terms of what these various combination methods achieve (e.g., whereas frequency contrasts are sensitive to changes in sample size, probability contrasts allow for comparability of contrast values across situations involving different focal-set sizes).

At the outset of this thesis, I discuss particulars of the present research that define its scope. I then address normative considerations that help shape the present framework. As noted earlier, documenting departures from normative prescriptions has been a key objective of the extant literature. Yet little detailed analysis of the appropriateness of these prescriptions has been undertaken (see Anderson & Sheu, 1995, for an insightful exception). To the extent that these prescriptions are inappropriate, claims of error may be exaggerated or simply incorrect. Following this discussion, I examine the literature on descriptive models of information integration in causal inference. Within this context, I outline some of the methodological and empirical problems that the rule-analytic approach faces and then move on to describe an alternative feature-analytic framework that I believe represents a resolution of these problems. Subsequently, I discuss five feature distinctions that are relevant to the substantive topic of information integration in causal inference. The descriptive importance of these distinctions is then evaluated with three experiments and with a reanalysis of data from five published experiments, following which conclusions are drawn.

SCOPE OF THE RESEARCH

Algebraic Rules as a Basis for Examining Feature Importance

Research on causal inference and contingency assessment is expanding rapidly in new directions. Even within the specific domain of information integration, many new theoretical approaches are being developed. For decades, portrayals of human causal inference were dominated by algebraic accounts that characterized people as intuitive statisticians who operate on frequency or probability information in order to form judgments (e.g., Peterson & Beach, 1967). More recently, however, new models and metaphors have emerged, being juxtaposed alongside the older ones. In particular, there is considerable interest in accounts based on principles of associative learning (e.g., Gluck & Bower, 1988; Shanks, 1991, 1993; Shanks & Dickinson, 1987). Although these accounts hold promise (for a review, see Allan, 1993), they are not without their limitations (e.g., see Melz, Cheng, Holyoak, & Waldmann, 1993). Indeed, within the rule-analytic tradition, both algebraic and associative models each have had their share of successes and failures, sometimes being "outperformed" by models emerging from yet other traditions (e.g., see Anderson & Sheu, 1995, Experiments 2 & 3).

While acknowledging the importance of these competing accounts, in the present research algebraic rules are used to examine feature importance. These rules are not viewed as models *per se* but rather as tools to explore feature importance. Given the ease with which algebraic rules exhibiting specific features can be constructed, they are useful for building a feature-analytic account of information integration in causal inference. However, the use of these rules should not be taken as a theoretical preference for algebraic models over associative ones. Indeed, the findings of this approach, theoretically, can support any type of model.

The Asymmetric Binary Case

A considerable amount of research has focused on how people integrate contingency information about two *binary* variables (for reviews, see Allan, 1993; Alloy & Tabachnik, 1984; Crocker, 1981; Shaklee, 1983). The focus on binary variables is especially relevant in the causal inference domain because often one is concerned with whether or not the occurrence or nonoccurrence of a particular event influences the occurrence or nonoccurrence of a subsequent event (Schustack & Sternberg, 1981). For instance, if someone wanted to determine whether Drug X caused nausea as a side effect, it would be informative to know how often nausea occurred when Drug X was either taken or not taken, as well as how often nausea did not occur when the drug was either taken or not taken. Similarly, in a social realm, if Richard wanted to test the notion that saying humorous things on a first date increased the likelihood of getting a second date with the same woman, he might consider how often he got a second date when he was either humorous or not particularly humorous, as well as how often he did not get a second date when he was either humorous or not particularly humorous.

More generally, Figure 1 shows the four possible event conjunctions derived by crossing an input variable (*I*) with an outcome variable (*O*). These conjunctions are typically labeled Cells A to D. In Cell A, the input and the outcome are both present (i.e., $I_p \cap O_p$, where the subscript *p* denotes presence). In Cell B, the input is present but the outcome is absent (i.e., $I_p \cap O_a$, where the subscript *a* denotes absence). In Cell C, the input is absent but the outcome is present (i.e., $I_a \cap O_p$). Finally, in Cell D, both the input and the outcome are absent (i.e., $I_a \cap O_a$). In this 2 × 2 contingency table, the letters *A* to *D* represent, respectively, the frequencies of the four conjunctions just described and *N* is the sum of these frequencies (i.e., the overall set size).

Variables whose levels reflect presence versus absence, such as those represented in Figure 1, are *asymmetric* because people tend to associate these variables with the positive state of event presence rather than with the negative state of event absence (Allan & Jenkins, 1980, 1983; Beyth-Marom, 1982). Here the terms *positive* and *negative* refer to what McGuire and McGuire (1992) called cognitive positivity (as compared with affective positivity). That is, these terms refer respectively to affirmational versus negational information, not the desirability or favorability of the information. The positive-negative asymmetry indicates that I and I_p are perceived as more similar than I and I_a (the same relation holds true for O). Asymmetric variables often can be restated in symmetric terms. For example, *red* versus *not red* can be restated as *red* versus *green*. One might view *red* as a more prototypic color than *green*, not to mention the fact that *green* may represent a small subset of *not red* cases. The point is that, like *red*, *green* is a positive example of color.

A connection between the symmetry of variables and the generation and testing of causal hypotheses also can be made. The asymmetric case is akin to evaluating a singular hypothesis (e.g., "X causes Y"), whereas the symmetric case is akin to evaluating two alternative hypotheses ("X causes Y" vs. "W causes Y"). Clearly, both types of variables are sometimes the focus of people's everyday causal thinking. However, given that people often consider singular hypotheses (for a review, see Kunda, 1990) especially when hypothesis-confirming data are available (Wason, 1960, 1968), asymmetric variables are relatively more likely to be the focus of everyday thinking.

Supporting this notion, Sanbonmatsu, Akimoto, and Biggs (1993) found that subjects overestimated the causal impact of a single focal cause even when other plausible causal candidates were available and subjects responded under accountability pressures (see Experiment 4). In line with other findings on debiasing (e.g., Lord, Lepper, & Preston, 1984), Sanbonmatsu et al. found that the causal impact attributed to a focal cause was attenuated only when subjects were explicitly told to consider other potential causal factors. I focus on the asymmetric binary case not only because of its ubiquity in everyday reasoning, but because only asymmetric variables allow for an examination of a central tenet of the present account; namely, that people weigh positive information more heavily than negative information in arriving at causal judgments.

NORMATIVE CONSIDERATIONS

With few exceptions (e.g., Arkes & Harkness, 1983; Inhelder & Piaget, 1958; Nisbett & Ross, 1980; Smedslund, 1963), the normative model of causal inference that has been adopted by researchers in the field is the delta rule,

$$\Delta \mathrm{P} = P(O_p|I_p) - P(O_p|I_a) = A/(A + B) - C/(C + D). \quad (1)$$

As shown above, 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. The ΔP rule is thus a measure of the one-way dependence of *O* on *I* (Allan, 1980) that assigns equal weight to the four cell frequencies. Equal cell weighting has thus been viewed by many researchers as a strict requirement for normative causal judgment. Anderson and Sheu (1995) have noted, however, that from a Bayesian perspective ΔP is not normative because of its set-size insensitivity (see also Allan, 1980). Moreover, Anderson and Sheu showed that when subjects are asked to evaluate a specific causal hypothesis (e.g., "X causes Y"), it is in fact normative to assign greater weight to *A* than to *D* because the former is likely to yield information of greater diagnosticity than the latter.

A more general point is that normative prescriptions need to be considered in relation to a judge's task objectives and concept definitions (Fischhoff & Beyth-Marom, 1983). If a judge's concept of causation focuses on what a logician would more precisely describe as a probabilistic criterion for assessing sufficient causation, then it makes no sense to describe the judge as erring in her information integration simply because she does not show evidence of testing necessary causation. One might argue prescriptively that judges *should* consider necessity as an important criterion in certain situations. However, the failure to do so represents a different type of error than the type committed by a judge who wants to weigh equally necessity and sufficiency criteria but who ends up focusing solely on information relevant to assessing sufficiency. The former reflects an error in criterion selection; the latter reflects an error in information integration.

To substantiate a claim of error in criterion selection, one must show that the chosen criteria are suboptimal for meeting the task objectives of the focal situation. In contrast, to show that a judge's information integration is error prone requires demonstrating that it is a suboptimal procedure for applying the judge's focal criteria (regardless of the appropriateness of these criteria). Given that most (if not all) of the research in this area has failed to probe for subjects' understandings of task objectives and of their subjective criteria for assessing causation (inferable from their definitions of causation), the various claims by researchers of nonnormative causal judgment on the part of subjects may be misleading. At the very least, such claims are difficult to interpret. Moreover, as other researchers (e.g., Arkes, 1991; Cosmides & Tooby, 1992; Friedrich, 1993) have argued, in practical real-world contexts normative modes of processing may exact costs that outweigh their benefits. Therefore, even if judges violate normative principles within an experimental context (where the imposed task objectives may have little resemblance to everyday task objectives), it is unclear that they are at a practical disadvantage for doing so. Indeed, the opposite may be the case.

To avoid confusion, I expand on Funder's (1987) error-mistake distinction. In the present scheme, a *bias* refers, in nonevaluative terms, to a deviation from a particular model or to the stronger weighting of one factor relative to another. An *error* refers to deviations from a normative model; clearly the model is viewed as appropriate if not optimal in a given experimental context and, consequently, error reflects suboptimal judgment or behavior at least in that context. A *mistake* refers to a judgment or behavior that is overly costly, suboptimal, or dysfunctional in a real-world situation, regardless of whether or not it constitutes an error vis á vis a given normative model. My main focus here is on biases in causal judgment. Some of the biases that I later predict do reflect error if ΔP is the assumed normative model. I suggest, however, that the same biases may also have pragmatic, if not adaptive, consequences in the real world.

DESCRIPTIVE MODELING WITHIN A RULE-ANALYTIC FRAMEWORK

A key objective of the literature on information integration in causal inference has been to identify the rule that best captures subjects' inferential responses on a variety of causal judgment tasks. This objective cuts across the various theoretical camps and is reflected in the data-analytic methods that have been employed. Before examining these methods, some basic features of a typical experiment should be described. First, subjects receive information about the contingency between two binary variables (e.g., taking or not taking drug X and getting or not getting side effect Y). Subjects are then asked to assess either the direction (e.g., Shaklee & Tucker, 1980), the magnitude (e.g., Allan & Jenkins, 1980; Schustack & Sternberg, 1981), or both the direction and the magnitude (e.g., Wasserman et al., 1993) of the inter-event relation, and this process is repeated over a set of stimuli that consists of different contingency data patterns. In some research (e.g., Anderson & Sheu, 1995, Experiment 1; Shaklee & Hall, 1983), subjects' reports of their judgment strategies also are collected.

Several experimental distinctions are noteworthy. Variables may be expressed in symmetric or asymmetric form, and task objectives may focus on assessing relation direction, relation magnitude, or both. Information is sometimes presented to subjects sequentially (over short time frames) using discrete-trial (e.g., Allan & Jenkins, 1980, 1983; Jenkins & Ward, 1965; Shanks, 1985) or continuous, free-operant (e.g., Chatlosh, Neunaber, & Wasserman, 1985; Wasserman et al., 1983) procedures. Alternatively, information is sometimes presented to subjects simultaneously, for example, as a set of contingency cases that subjects are free to sort (e.g., Smedslund, 1963, Experiment 2) or as summarized data grouped by cell (e.g., Kao & Wasserman, 1993, Experiment 2; Ward & Jenkins, 1965). The questions posed to subjects also have varied, for example, focusing on subjects' *control* of an outcome variable (e.g., Wasserman et al., 1993), the *effectiveness* of one event on promoting another (e.g., Kao & Wasserman, 1993), or the *probability* that one event causes another (e.g., Schustack & Sternberg, 1981). Here, subjects' responses to these various questions are generically termed *causal ratings* and are viewed as behavioral manifestations of an implicit causal inference process.

Rule-Analytic Methods

The two primary types of methods for modeling causal ratings that have been employed in the rule-analytic tradition are standard correlational methods and rule analysis. These methods (which address the question of what rule is best) and their limitations are discussed next.

Standard Correlational Methods

Correlational methods have been used in two ways that differ primarily in terms of the unit of correlational analysis. In *strategy classification*, each subject's ratings are correlated with each rule under consideration. A criterion for nonexclusive classification is simply to tally the number of subjects whose ratings are reliably correlated with a given rule at either a given Type I error rate (e.g., $\alpha = .05$) or beyond a given correlational magnitude (e.g., Allan & Jenkins, 1983, used a criterion of $r \geq .70$). The descriptive utility of a rule is thus a function of the proportion of the sample accounted for by that criterion. A criterion for exclusive classification is to group subjects according to the rule that correlates the strongest with their ratings (e.g., Allan & Jenkins, 1980). In *model fitting*, ratings first are averaged across subjects and then regressed on each focal rule (e.g., Anderson & Sheu, 1995; Schustack & Sternberg, 1981). The rules are then evaluated in terms of explained variance, as well as the weights assigned to rule components.

Limitations of Standard Correlational Methods

An advantage of strategy classification is that it focuses on the individual as the unit of correlational analysis. As Anderson (1981) noted, this is the appropriate level of analysis for basic research on information integration because it is sufficient to provide information about both individual differences and nomothetic trends. Nevertheless, strategy classification is lacking because it relies on absolute criteria that are arbitrary. For instance, in using a criterion of $r \geq .70$, a correlation of .69 would be distinguished from one of .70, yet the latter value would not be distinguished from a correlation of .96! This example reveals how imposing absolute criteria can result in a loss of important metric information. Strategy classification by an exclusive criterion faces a related problem: Although grouping is not based on an absolute criterion (i.e., it is based on the highest correlation coefficient), it may still be problematic. For example, if the "best rule" correlates .90 and the second-best rule correlates .89 with a subject's ratings, clearly this should not be viewed as a meaningful difference.

Although model fitting obviates the need to invoke absolute criteria for classification, it suffers from other problems. First, it is typically applied to group-averaged data which, as noted earlier, is not the proper level of analysis. Second, although significance testing can be performed in order to determine whether two models differ in their predictive utilities, such significance testing is seldom done. More commonly, rules are judged on the basis of the highest R^2 value. Third, model fitting allows parameters to be empirically determined. The predictive utilities of "good models," however, may be significantly attenuated in attempts at cross-validation. More importantly, the focus on the predictive utilities of various rules steers attention away from the theoretical objective of discovering *design principles* (Stone & Van Orden, 1994)—principles that relate model behavior to observable human behavior.

Rule Analysis

The realization that not only are distinct rules collinear but sometimes perfectly correlated (Allan, 1980) led Shaklee and her colleagues (Shaklee, 1983; Shaklee & Elek, 1988; Shaklee & Hall, 1983; Shaklee & Mims, 1981, 1982; Shaklee & Tucker, 1980; Shaklee & Wasserman, 1986; see also Arkes & Harkness, 1983) to adopt an alternative method for rule discrimination called *rule analysis*. Essentially, rule analysis entails using a set of stimuli that is constructed to discriminate between rules. In a case in which three rules are being evaluated, for instance, some stimuli lead to predictions in the same causal direction for all three rules being assessed, other stimuli discriminate Rule 1 from Rules 2 and 3, and some stimuli discriminate between all three rules. If the relation direction indicated by a judge corresponds, for example, two or three times out of three, the judge is deemed to have met the criterion for classification under a given rule category.

Limitations of Rule Analysis

As others (e.g., Busemeyer, 1991; Kao & Wasserman, 1993) have noted, rule analysis suffers from many shortcomings. First, like strategy classification, it relies on absolute criteria.

Second, the sample of problems on which the more complex rules are assessed is typically very small. Third, rule analysis does not take into consideration subjects' assessments of the magnitude of causal relations. And, even if rule analysis is adapted to take into account magnitudes, the cutoff criteria for strategy classification are still arbitrary (e.g., see Kao & Wasserman, 1993). The main advantage of rule analysis is that it forces researchers to think about how to construct stimulus sets that allow for rule discrimination. However, this effort can also be taken with correlational methods that provide indices of accuracy that need not be arbitrary and that take assessments of relation magnitude into consideration. One might even argue that rule analysis obscures poor rule discriminability by establishing a procedure that loses "incriminating" metric information about rule multicollinearity. At any rate, rule analysis, like the correlational method, focuses on assessing rule superiority rather than feature importance.

Self Reports of Strategy Use

Self reports of strategy use have been used in addition to rule-analytic methods to classify subjects' strategies. Accordingly, subjects are asked to describe the thinking they used to arrive at their causal ratings. Given that people's explicit understanding and reporting of their own implicit cognitive processes are often misguided (for reviews, see Abelson & Levi, 1985; Ericsson & Simon, 1980; Nisbett & Wilson, 1977; Slovic, Lichtenstein, & Fischhoff, 1985; Wilson, 1994; see also Wilson & Brekke, 1994), self-report data alone are of questionable value in formulating a judgment model. However, self reports may be useful in identifying subgroups of strategy users when these reports are provided following a judgment task (Anderson & Sheu, 1995).

Rule-Analytic Findings: Still No Winner?

Over the past 30 years, researchers have explored the rules that people use to integrate contingency information. The preceding discussion highlighted some of the limitations of rule-analytic methods employed in that research. A more serious problem facing the rule-analytic approach, however, is its failure to identify a stable empirical ranking of rule superiority even among a relatively circumscribed set of plausible models. Although a detailed account of the successes and failures of each evaluated rule runs counter to the feature-analytic objectives of the present research, a brief review of these findings, as well as a description of the most frequently evaluated rules, is in order.

Given the (perhaps unwarranted) popularity of ΔP as a normative model, much research has focused on the viability of this rule as a descriptive model (for recent reviews, see Allan, 1993; Kao & Wasserman, 1993; Wasserman, 1990). Most of the algebraic alternatives have been labeled "linear combination heuristics" (Cheng & Novick, 1992) and have been selected because of their deviations from the standard normative model. Notably, each of these alternative strategies relies on frequency rather than probability information, and some strategies do not utilize data from each of the four cell frequencies (thus violating the equal-cell-weighting requirement).

The linear heuristics that have received the most attention include the Cell-A strategy, the confirming-cases strategy, the *A* minus *B* strategy (*A–B*), and the difference-in-the-sums-of-diagonals strategy (ΔD). The Cell-A strategy involves checking to see whether *A* > *B*, *C*, *D*, and if so, a positive contingency between *I* and *O* is judged. A variant of the Cell-A strategy, termed by Allan and Jenkins (1983) the *success-counting* rule, is to assess contingency strictly on the basis of *A*. The confirming-cases strategy involves summing either *A* and *D* (confirming a facilitative relation) or *B* and *C* (confirming an inhibitory relation). *A–B* involves subtracting *B* from *A*. ΔD involves summing *A* and *D* and subtracting the sum of *B* and *C*. For the latter two strategies, the relation is judged positive, negative, or noncontingent for positive, negative, and zero values, respectively, with judged magnitude being a positive function of the absolute predicted value.

Across studies, the Cell-A strategy (e.g., Jenkins & Ward, 1965; Smedslund, 1963), the confirming-cases strategy (e.g., Ward & Jenkins, 1965), *A–B* (e.g., Wasserman, Dorner, & Kao, 1990), ΔD (e.g., Allan & Jenkins, 1983; Shaklee & Tucker, 1980), as well as ΔP (e.g., Anderson & Sheu, 1995, Experiment 1; Wasserman et al., 1983), at some point or another, have all been found to be used most frequently or to be most highly correlated with subjects' causal ratings. Furthermore, simply tallying the number of studies supporting each of these rules is inconclusive because some rules are evaluated more often than others. In short, as Allan (1993, p. 438) stated in a recent review of the literature, "the quest for a rule to describe human judgments of the contingency between binary variables has not yet yielded a satisfactory solution."

What is the basis for the general inconclusiveness of this rule-analytic quest? Differences in experimental features across studies surely account for some of the variance in rule rankings. For instance, the format of data presentation is likely to influence strategy use (Beyth-Marom, 1982; Crocker, 1981). When data are shown in summary format, subjects tend to adopt more complex strategies such as ΔP (e.g., Kao & Wasserman, 1993; Ward & Jenkins, 1965), presumably because of a reduction in memory demands under these conditions (Yates & Curley, 1986). Several studies also have used instructions that one might reasonably expect would lead subjects to respond in a particular way (for reviews, see Beyth-Marom, 1982; Cheng & Novick, 1992; Crocker, 1981). For example, Ward and Jenkins (1965) provided their subjects with a definition of control that explicitly defined Cells A and D as confirming cases. Accordingly, they found that subjects based their judgments on the sum of *A* and *D*. As others (e.g., Hilton, 1990; Krosnick, Li, & Lehman, 1990; McGill, 1989) have noted, given that subjects treat the experimental context as a social and communicative situation, subjects respond to subtleties in instructional wording and experimental detail, and they are likely to expect that everyday conversational norms (Grice, 1975) apply in that context.

Aside from documenting experimental features that are likely to bias subjects toward using particular strategies, however, little advancement in understanding the basis for rule-analytic inconsistencies has been made. In the next section, I argue that rule-analytic consistency may be impossible, even in theory, because any particular rule consists of multiple features, some of which may consistently be important in accounting for rule viability and some of which may be of lesser importance. In so doing, I lay the foundation for a new *feature*-analytic framework for studying information integration strategies in causal inference.

A FEATURE-ANALYTIC FRAMEWORK

A Theoretical Challenge to Rule-Analytic Research

The preceding section highlighted some of the methodological limitations and empirical inconsistencies of the rule-analytic approach. Yet the objective of finding the "best" descriptive rule can also be challenged on theoretical grounds. To begin with, it is important to question what constitutes a rule. Descriptive rules may be viewed as a formal instantiation of a conjunction of features. A *feature* is an attribute that is shared by a given set of rules.[1] Although features can be quantitative (e.g., number of operations in an algorithm), the focus here is on qualitative features which I describe in more detail later on.

From a feature-analytical perspective, to demonstrate that one rule is associated more strongly with people's responses than another rule is essentially to demonstrate the superiority of one set of features over another set of features in characterizing those responses. Thus, to assume that a "best rule" can be found thus presupposes that there exists a set of features that, when combined, consistently outperforms other sets of features. It is plausible, however, that whereas some features are especially important in defining rule viability (reflecting more or less invariant human information processing tendencies), other features are of lesser importance and are more easily influenced by contextual factors. If so, one would expect the set of rules that possesses the consistently important features to outperform the set of rules that does not possess those features. But, even in theory, one might expect that, within the set of "high performance" rules, different rules would dominate on different occasions depending on particulars of the information-processing environment. Therefore, this would preclude the emergence of a single best rule (and frustrate the search for one).

Note that this account of the inconsistency in rule ranking across studies cannot simply be treated as an artifact stemming from experimental biases such as leading question wording. It is plausible that the nature of the contingency data *per se* might lead to different rules being favored. That is, aspects of the data or the environment in which the data are observed may make some methods for integrating information simpler or more salient than other methods. Given that real-world contexts vary in the types of information that are available or that are salient to observers, context-dependent variation in inferential-strategy use is a non-trivial issue. Such variation may indicate that people are equipped to use different processing routes in order to arrive at the same type of judgment.

Supporting this assertion, Anderson and Sheu (1995) found that, using a discrete paradigm in which data on the pairing of cause and effect conjunctions were readily available, on average, subjects adopted a weighted ΔP strategy. However, using a paradigm in which causes and effects occur continuously in time (and in which the subject controls the rate at which the cause occurs), Anderson and Sheu found that subjects' causal ratings were a function of rates (i.e., probabilities per unit of time) rather than probabilities. The investigators suggested that causal judgments are a function of perceptually salient variables in the environment.[2] Context-dependent accounts, such as Anderson and Sheu's account, imply that the objective of finding a single best rule is theoretically misguided.

Representations of Similarity in Descriptive Modeling

The present feature-analytic framework draws on Tversky's (1977) set-theoretical (feature-based) account of similarity judgments. Tversky argued that similarity judgments violate assumptions that characterize metric representations of similarity. His theory proposes instead that objects are described by sets of features, and that the perceived similarity between objects results from a feature-matching process. Tversky's descriptive theory favors a set-theoretical representation over a metric one. The present framework, however, uses both types of representation. Feature representations are used to characterize both judgment processes and rule structure. Metric representations are used to characterize the outcomes of judgment processes and rules. I now expand on these concepts.

The search for the best rule has much to do with judgments of similarity. First, researchers seek out rules whose predictions are highly similar to subjects' ratings. These rule-judge comparisons address the issue of rule accuracy. Second, researchers are interested in showing not only that a rule has a certain level of accuracy but that it surpasses the accuracy levels of other rules. This objective of rule discrimination involves both inter-rule comparisons and inter-rule-judge contrasts. Thus, there are three types of similarity estimates that are relevant to assessing accuracy and discrimination.

The standard correlational method provides metric estimates (e.g., r^2) of both accuracy and discrimination. Consider two rules (R1 and R2) that are being evaluated as descriptive models of a judge (J1). The accuracies of R1 and R2 are given by the correlations of their respective predictions with J1's ratings. The difference between these two correlations (i.e., the inter-rule-judge contrast) is an estimate of discrimination, the magnitude of which will be inversely related to the correlation between R1 and R2 (i.e., this inter-rule comparison assesses rule collinearity). An important feature of these correlational estimates of accuracy and discrimination is that they are based on the rules' *predictions* and the judge's *ratings*. That is, they are based on outcomes, irrespective of the generating processes. Thus, the accuracies of two different rules that yield the same predictions for a given data set are estimated to be identical, despite their differing computational structures. For example, as Allan (1980) noted, if $P(I_p) = P(O_p) = .5$, then ΔP, ΔD, and *A–B* all yield identical predictions and are thus indiscriminable in correlational terms.

In contrast to the correlational method that focuses on outcomes, consider a set-theoretic representation of inter-rule similarity that focuses on process structure. Rules may be distinguished from each other on the basis of the overlapping and distinct features that may be attributed to them. Assessing rule similarity in terms of features is advantageous because features can be defined and attributed to rules independent of the specific prediction patterns of those rules for a given set of stimuli. Referring to the previous example, even if the base rates for levels of *I* and *O* are equal, ΔP, ΔD, and *A–B* can be discriminated from one another on the basis of their distinct features. For example, whereas ΔP combines probabilities the latter two rules combine frequencies. Or, whereas the former two rules integrate data from all four cells, *A–B* relies on only Cells A and B.

The power of analyzing features, however, becomes most apparent when correlational estimates of accuracy are jointly considered. Whereas two rules may be contrasted in terms of their overlapping and distinct features, a corresponding analysis of rule-judge similarity cannot be undertaken. This is because it is the judgment process *per se* that ultimately is unknown. If the features characterizing that process were known, then modeling it would be straightforward. Here the connection between correlational and feature-based representations of similarity becomes useful both methodologically and conceptually.

On the one hand, correlational estimates of rule accuracy are particularly useful because we do not know the features characterizing people's information integration strategies. We thus assess accuracy based on what we do have; namely, causal ratings generated from an unknown judgment process. But what we ultimately want to know is something about the features that characterize the judgment process itself. On the other hand, we have at our disposal multiple rules to model that process. Moreover, each of these rules can be defined in terms of various features. We now can ask, what are the features associated with a particular rule that contribute to its accuracy?

Feature Analysis

The preceding question does not ask what rule is best, but rather what features are important in describing the causal inference process. In contrast to rule-analytic techniques, I introduce a technique called *feature analysis* that is aimed at addressing the latter question. The underlying logic of feature analysis is as follows: If a rule is represented by a set of features, then evaluating the importance of those features can be accomplished by (a) feature coding various rules, any two of which differ in terms of at least some features and (b) feature testing, that is, partitioning the variance in accuracy in terms of the coded features. Statistically, then, feature analysis entails adding one or more additional steps to the standard correlational method, the last of which entails analysis of variance. Analyzing variance in correlational estimates of accuracy is a powerful technique because it allows for the importance of both specific features and feature interactions to be assessed.

Let me make explicit a notion that was implied in the preceding paragraphs. As just noted, the objective of feature analysis is *not* to find a better way to identify the best rule, but rather to identify which rule features best capture the essence of human inferential strategies. Single features do not, however, give rise to contingency estimates as do single rules. Because features are noncomputational entities, they must be studied indirectly through the rules to which they may be attributed. Rule predictions and judges' causal ratings are thus treated as intermediaries between features of rules and features of unknown judgment processes. Therefore, the feature-analytic approach does not reject all aspects of the rule-analytic approach. Whereas the two approaches differ in their objectives, both are analytic. In effect, the feature-analytic approach rests upon the rule-analytic framework but modifies it in a critical manner. Keep in mind that from a feature analytic perspective, rules do not necessarily provide models of either conscious or implicit information integration strategies. As noted earlier, rules model the output of information integration processes, whereas inferences about the attributes of the processes themselves are made on the basis of feature-analytic findings.

Accuracy: Predictive Utility or Model Viability?

The logic underlying feature analysis is that, if a given feature is important in representing how people integrate information in causal inference then, *ceteris paribus*, rules that possess that feature should have greater accuracy than rules that do not possess the same feature. In one sense, accuracy connotes predictive utility. A direct measure of a rule's predictive utility is r^2—the proportion of variance in a subject's causal ratings that is explained by the rule. One problem with analyzing r^2 values, however, is that they do not take into consideration the *direction* of the relation between a judge's ratings and a rule's predictions. Because rules are not only viewed as predictors but also as *models* of human inferential judgment, it is important to distinguish between positive and negative correlations. That is, a rule's predictions should be positively correlated with subjects' ratings if it is to be coded as a *viable* model. Consider the situation in which the predictions of a given rule are correlated with two different raters' responses with equal magnitude but in opposite directions. From a theoretical perspective, these two predictive effects should cancel out one another rather than be added together. Thus, correlation coefficients are a more appropriate estimate of accuracy than squared coefficients. Moreover, because these coefficients represent feature-analytic data, they are Fisher transformed (Fisher, 1921) from r to r' in order to improve their distributional properties. I refer to these r' values as *viability estimates*.

Level of Correlational Analysis

In line with the tenets of information integration theory (Anderson, 1981), feature analysis is conducted at an individual-level of correlational analysis. At this level, and in contrast to model fitting, each subject's ratings are first correlated with each of the rules. Only then are these rule-judge correlations averaged across subjects.

A FEATURE-BASED ACCOUNT OF CAUSAL INFERENCE

Distinctions Concerning the Type of Information that is Integrated

I will begin by discussing three feature distinctions that concern the type of contingency information that judges integrate. These include the contrastive-cumulative, positive-negative, and input-outcome distinctions. After defining each distinction, propositions about features of the information integration process in causal inference are made in terms of these distinctions and their interactions, and a fourth distinction between sufficiency and necessity testing is introduced

The Contrastive-Cumulative (C-C) Distinction

Most of the rules that have been examined are what I term *contrastive*. These rules entail a contrast between different types of information. Evidence for the viability of contrastive rules in modeling judges' responses would suggest that judges may be attempting to calculate tradeoffs between conflicting types of evidence. In the domain of causal inference, these tradeoffs are between information that confirms a focal relation and information that disconfirms that relation. From a logico-statistical perspective, Cells A and D provide confirmation of a facilitative causal relation and Cells B and C provide disconfirmation of such a relation. In terms of inhibitory causality, the implications of these cells are reversed. Because the confirmatory nature of information depends on the direction of the relation in question, I refer to *A* and *D* as confirmatory cases of a facilitative relation (CFAC cases) and *B* and *C* as confirmatory cases of an inhibitory relation (CINH cases). Thus, contrasts generally are defined in terms of tradeoffs between CFAC cases and CINH cases. For example, *A–B* contrasts a subset of CFAC cases (*A*) with a subset of CINH cases (*B*); similarly, ΔD contrasts the sum of all CFAC cases with the sum of all CINH cases in a focal set.

In contrast, *cumulative* rules pool information from distinct cells as a summary of either marginal frequencies or probabilities, CFAC cases, or CINH cases. Marginal rules entail calculating the marginal frequency or probability of either I_p, I_a, O_p, or O_a (there is no need to distinguish between event-present or event-absent states because the two marginal probabilities always sum to the constant 1). Cumulative rules that focus either on summing the frequency of CFAC cases or

CINH cases are identical to what was described earlier as a confirming-cases strategy.[3] The important feature shared by cumulative rules, in contrast to contrastive rules, is that no tradeoff between conflicting types of evidence is necessarily implied.

The Positive-Negative (P-N) Distinction

As noted earlier, people exhibit a preference for information about event presence rather than event absence (Beyth-Marom, 1982; Fazio, Sherman, & Herr, 1982; Halberstadt & Kareev, 1995; Klayman & Ha, 1987; Nisbett & Ross, 1980; Wason, 1959). In this context, positive and negative information is defined by event presence (I_p or O_p) and event absence (I_a or O_a), respectively. Klayman and Ha (1987) drew a distinction between judgment strategies that rely on positive information (+test strategies) and those that rely on negative information (–test strategies). Specifically, if part of a rule or strategy involves contrasting *A* and *B* (viz., data from the two I_p cells) or *A* and *C* (viz., data from the two O_p cells), that part is referred to as a +test. Contrasts involving *D* and *B* (viz., data from the two O_a cells) or *D* and *C* (viz., data from the two I_a cells) are termed –tests. For example, in terms of the P-N distinction, ΔP can be defined as the combination of one +test (i.e., $P[O_p|I_p]$, focusing on I_p cases) and one –test (i.e., $P[O_p|I_a]$, focusing on I_a cases). Although Klayman and Ha defined the P-N distinction at an event level, it can be extended to event conjunctions. Indeed, the only element that distinguishes +tests from –tests is inclusion of *A* in the former case and *D* in the latter case. The positivity-negativity of Cells A to D is +/+, +/–, –/+, and –/– for variables *I* and *O*, respectively. Therefore, in terms of positivity, Cell A > Cell B = Cell C > Cell D.

The Input-Outcome (I-O) Distinction

The I-O distinction distinguishes between contrasts that take place within either levels of the input variable (Itests) or levels of the outcome variable (Otests). Contrasts involving *A* and *B* (viz., data within I_p) or *C* and *D* (viz., data within I_a) are Itests. Contrasts involving *A* and *C* (viz., data within O_p) or *B* and *D* (viz., data within O_a) are Otests. For example, ΔP can be defined more precisely as a contrast between a +Itest and a –Itest. Unlike the P-N distinction, the present one cannot be defined at the level of event conjunctions because each cell involves a level of *I and* a level of *O*.

The I-O distinction seemingly bears a close resemblance to Klayman and Ha's (1987) distinction between hypothesis tests (Htests) and target tests (Ttests). Klayman and Ha were concerned with the 2 × 2 case, but focused primarily on hypothesis testing and rule discovery (e.g., Wason, 1960, 1968; Wason & Johnson-Laird, 1972). In Klayman and Ha's account, the row variable represents cases that either conform to (top row) or do not conform to (bottom row) a hypothesized rule that a subject has generated. The column variable represents cases that either conform to (left column) or do not conform to (right column) a target rule that the subject would try to infer based on accurate feedback. In the present account, *I* is always the row variable and *O* is always the column variable. So, in terms of causal inference, Itests are identical to Htests and Otests are identical to Ttests given that a judge is considering the hypothesized cause of a target outcome. However, this correspondence is unnecessary because a judge may instead hypothesize the effect of a target input. In the latter case, Itests and Ttests correspond, as do Otests and Htests. Moreover, the causal relation between two events can be evaluated without ascribing target status to one particular event. In such cases, as in the experiments reported later, there is no basis for coding rules as Htests or Ttests.

Propositions Concerning the Type of Information that is Integrated

Subjective Cell Importance Findings: An Empirical Basis for Prediction

Recent research (e.g., Anderson & Sheu, 1995; Kao & Wasserman, 1993; Levin, Wasserman, & Kao, 1993; Wasserman et al., 1990) has implemented powerful techniques (e.g., selective pairwise stimulus comparisons) for assessing the subjective importance or weights that judges assign to particular cells in causal judgment tasks. The literature on subjective cell importance (SCI; e.g., Anderson & Sheu, 1995; Crocker, 1982; Kao & Wasserman, 1993; Levin et al., 1993, Experiment 1; Schustack & Sternberg, 1981) clearly indicates that, on average, people weigh $A > B > C > D$. Wasserman et al. (1990), for example, found that this general pattern held when (a) subjects were asked to indicate explicitly what types of information are important for contingency judgment, (b) subjects were asked to explicitly rate SCI on a continuous scale, and (c) SCI was assessed implicitly via subjects' ratings on a contingency assessment task. Moreover, on average, judges weigh A and D (i.e., CFAC cases) positively and they weigh B and C (i.e., CINH cases) negatively, indicating that their interpretations of the evidential implications of the four cells are consistent with logical accounts (for a meta-analysis demonstrating this finding, see Lipe, 1990).

The fact that this SCI pattern obtains across many studies by independent researchers using different experimental designs attests to its robustness. Indeed, in an area of research governed by empirical inconsistencies in rule rankings, the stability of the relative weighting of cells is striking. Nevertheless, accounts of why these findings obtain have been lacking. The interplay between these findings and the current feature-analytic approach is reciprocal: On the one hand, SCI findings suggest what features might be favored. On the other hand, by relating these empirical findings to definable features of a judgment process, theoretical advancements in understanding their origin can be made. The propositions subsequently discussed are intended to build on each other, resulting in a theoretical account that is tied to well-established findings in causal-inference and related domains.

Proposition 1: Judges Consider Evidential Tradeoffs

The SCI findings indicate that, although judges deviate from the equal-cell-weighting requirement, they do (on average) consider evidential tradeoffs. For example, if their judgment processes were better characterized by a (cumulative) confirming-cases strategy in which judges simply considered f(CFAC cases) than by a constrastive strategy, then perhaps we would expect judges to weigh $A \cong D > B \cong C$. The simple fact that the mean absolute weight assigned to A and B is consistently greater than the mean absolute weight assigned to either A and D *or* B and C indicates that, on average, a +Itest (which is contrastive) will outperform both of the confirming-cases tests. Of course, this does not mean that *any* contrastive test will outperform a cumulative test. In general though, a contrastive test should outperform a cumulative test in terms of its rule viability when both tests are matched in terms of the type of data on which they operate.

There is reason to expect that judges will tend toward the use of contrastive strategies. Contrasting types of information that differ in their implications for a given hypothesis is an essential feature of diagnostic hypothesis testing and hypothesis revision. Without considering such contrasts, a judge would have no opportunity for disconfirming a hypothesis or an existing belief. Indeed, in the domain of social information gathering and hypothesis testing, researchers (e.g., Bassok & Trope, 1984; Skov & Sherman, 1986; Trope & Bassok, 1982, 1983) have shown that judges preferentially seek information of greater diagnosticity. That is, judges seek information that maximizes the contrast between a hypothesis and its alternative. Moreover, subjects in these studies preferred diagnostic information to information that was more likely to confirm their target hypothesis.

The tendency to seek diagnostic information, and thus to consider evidential tradeoffs, is adaptive. However, the fact that judges seek diagnostic information does not mean that they are nonselective in the types of errors that they try to minimize (Friedrich, 1993). For example, whereas judges tend to narrow overly general hypotheses, they often fail to broaden overly restrictive hypotheses, posing problems especially in situations involving hypotheses that already are too narrow (Klayman & Ha, 1989). Nevertheless, although judges may be biased towards revising hypotheses in some regards and not others, on the whole, the fact that they do engage in some types of hypothesis revision indicates that they are using contrastive strategies.

Findings by Wasserman et al. (1990, Experiment 1) support the contention that judges consider contrasts to be an important part of the judgment process. For example, in one condition the investigators asked judges to check off which of the four cells were necessary to consider in order to assess contingency between two binary variables (viz., whether or not a drug was administered and whether or not a rash subsequently developed). Nearly half (47%) of the subjects indicated that all four cells were necessary for judgment. An additional 42% checked either Cells A and B, Cells A and C, or Cells A, B, and C. In contrast, only 6% checked only Cells A and D or Cells B and C, and only 4% checked only one cell. Bearing in mind that, on average, subjects weigh Cells A and D positively and Cells B and C negatively, these findings indicate that subjects understand the need to contrast conflicting types of information.

Proposition 2: Judges Display a Stable +test Bias

The fact that judges tend to weigh $A > B > C > D$ indicates that they weigh positive information more heavily than negative information because Cell A is +/+ whereas Cell D is –/–. That is, judges are biased toward +testing. If so, the viabilities of +tests should greatly exceed those of –tests. In addition to the preceding weighting inequality, there are several other areas of research that cohere with this prediction. For instance, people tend to learn concepts more efficiently from positive information than from negative information even when the informativeness of both sources are kept equal (Hovland, 1952; Hovland & Weiss, 1953). Wason (1959) demonstrated this effect, showing that people are quicker at processing information about affirmative statements (e.g., "there is X") than about negative statements (e.g., "there is not X"), regardless of whether the statements were true or false.

Not only are people faster at processing positive information, they are more likely to learn to discriminate between, or predict, positive and negative states (e.g., a light being on or off) when a single diagnostic cue is associated with the positive state than when it is associated with the negative state. In the former (and more effective) case, the positive state is signaled by the presence of the cue, whereas in the latter case the positive state is signaled by the absence of the cue. The fact that learning is better in the former case suggests that discrimination is better achieved by positive information than by negative information. This feature-positive effect, as it has been termed, has been demonstrated in both humans (Newman, Wolff, & Hearst, 1980) and many nonhuman species (Jenkins & Sainsbury, 1970; for a review, see Hearst, 1978).

Evidence also suggests that hypothesis testing is governed by a +test strategy. For example, Klayman and Ha (1989) categorized the types of test strategies used by their subjects in a rule-discovery task. Out of the 18 hypothesis-testing opportunities that subjects were given, on average, about 12 consisted of +tests, less than 2 consisted of –tests, and about 4 consisted of alternative tests; that is, tests of specific nontarget hypotheses that were pitted against the current hypothesis (there are no alternative tests in the present account which examine the situation involving only one antecedent variable and one consequent variable). Therefore, +tests were twice as common as the other two types of tests combined.

In addition to empirical evidence, there are also some plausible accounts of why people tend to be biased toward positive information. One account is based on informativeness. As Wason argued, it is more informative to learn what *is* the case than what is *not* the case. Indeed, learning that an event does not belong to a given category serves primarily as an opportunity to search for, or to construct, other categories that do encompass the event. Consistent with this interpretation, a *not* symbol sometimes refers to the absence of an event or category (as in Figure 1) but other times is treated as an exclusion symbol that refers to anything but the event being negated. Therefore, *not* may often serve to refine a set of possible features rather than emphasize features that are known to be absent.

Another account that is related to the previous one is based on the ambiguity of negative information. Compared with positive information, negative information is considerably more ambiguous and this ambiguity may cause difficulties in defining a target sample space. For example, if the label *not red* is used as an exclusionary term, does it refer to events defined by a subset of other colors, events defined by all possible colors other than red, or events that are not even defined in terms of color? Indeed, Einhorn and Hogarth (1986) suggested that one reason why people tend to attach little importance to Cell D is that, in the absence of an assumed reference population, it defines a vague set of instances. Given this lack of specificity in negative information, it is perhaps not surprising that language itself tends to favor affirmatively-phrased statements (Clark & Clark, 1977).

From a subjectivist perspective on diagnosticity, there is also reason to favor +testing over −testing. As Anderson and Sheu (1995) detailed, the joint occurrence of the hypothesized cause and effect is more diagnostic than their joint absence because, intuitively, $P(O_p|I_p)$ should be much greater than $P(O_p)$; yet $P(O_p|I_a)$ should be much closer to $P(O_p)$. Therefore, the latter conditional probability (which utilizes C and D) should be less diagnostic than the former (which utilizes A and B). Indeed, if $P(O_p|I_a) = P(O_p)$, then D is nondiagnostic. In sum, considerable evidence (and some plausible rationales) point toward a +test bias.

Proposition 3: Judges Are Biased Toward Itests Given +testing and Otests Given –testing

Given the SCI findings, the P-N and I-O distinctions should interact. To see why, consider the cell frequencies used by each type of test: +Itests contrast *A* and *B*, –Itests contrast *C* and *D*, +Otests contrast *A* and *C*, and –Otests contrast *B* and *D*. If judges weigh $A > B > C > D$, one would expect that, in terms of rule viability, +Itests > +Otests > –Otests > –Itests. Thus, under +test conditions, Itests should outperform Otests, but under –test conditions, Otests should outperform Itests.

The previous findings by Klayman and Ha (1989) offer a clue as to why this may be so. They found that the frequency of +testing was roughly three times greater than the testing of specific alternatives. This suggests that +test biases stem from factors other than the mere ambiguity of negative information. That is, judges favor +testing even when alternatives can be well defined. As noted earlier, these findings indicate that judges have a preference for narrowing overgeneralized hypotheses rather than broadening overly narrow ones. If a hypothesized cause is overly general sometimes it will be present even when the target outcome is absent. In causal terms, it will be *insufficient* to produce the target outcome.[4]

Now consider what broadening an overly narrow hypothesis entails in the causal inference domain. If a hypothesized cause is too narrow it will sometimes be absent when the target outcome is present. In causal terms, it will not be *necessary* to produce the outcome. This interpretation suggests that, in terms of a sufficiency-necessity (S-N) distinction, judges have a bias for testing the sufficiency (Stesting) rather than necessity (Ntesting) of a cause. In many situations it is pragmatic to weigh a sufficiency criterion more heavily than a necessity criterion. As Klayman and Ha (1989, p. 603) put it, "We don't mind passing up some potentially acceptable cars if we can avoid buying a lemon." In other words, in many real-world contexts, judges are willing to tolerate necessity violations in order to minimize sufficiency violations. As Friedrich (1993) suggested, the costliness of sufficiency violations often exceeds that of necessity violations. Thus, judges operate under the implicit guideline, "if it ain't broke, don't fix it." Translated into causal terms, if a hypothesized cause is found to be sufficient, don't waste time searching for other possible causes.

The SCI findings, suggestive of a P-N × I-O interaction, can now be interpreted in light of the proposed +test and Stest biases. Consider first the +test case: +Itests contrast *A* and *B* thus providing information about sufficiency regardless of whether the causal relation is hypothesized (or perceived) to be facilitative or inhibitory. Under a "facilitative set," Cell B is the "sufficiency cell" or Scell; under an inhibitory set, Cell A is the Scell. In contrast, +Otests provide information about necessity under a facilitative set (Cell C is the Ncell) and information about sufficiency under an inhibitory set (Cell A is the Scell). Therefore, unlike +Otesting, +Itesting consistently provides a basis for Stesting. Now consider the less-favored –test case: –Itests contrast *C* and *D* thus providing information about necessity under both facilitative (Cell C is the Ncell) and inhibitory (Cell D is the Ncell) sets. In contrast, –Otests provide information about sufficiency under a facilitative set and they provide information about necessity under an inhibitory set. Therefore, unlike –Itesting, –Otesting sometimes provides information relevant to Stesting.

Proposition 4: Judges Display a Qualified Itest Bias

Some researchers (e.g., Klayman & Ha, 1987; Tversky & Kahneman, 1980) have suggested that people may exhibit a bias toward Itesting because only Itests are consistent with the experience of causality flowing forwards in time. Thus, Itests should "make more sense" to people. Because Itests operate on data from a distinct value of *I*, they imply questions like, "given that Antecedent X occurred (or did not occur), how likely is Consequent Y to occur (or not to occur)?" In contrast, Otests make the antecedent conditional on the consequent, resulting in questions like, "given that Consequent Y occurred (or did not occur), how likely was Antecedent X to occur (or not to occur)? Although the latter is a valid diagnostic question, it violates the forward temporal direction inherent in people's perception of causality. Providing some indication that people may be biased toward Itesting, Tversky and Kahneman (1980) demonstrated that judges are more confident in making predictions about *O* on the basis of information about *I* than vice versa. As the previous predictions indicate, however, an I-O main effect would likely be qualified by a P-N × I-O interaction. This suggests that an Itest bias in causal inference is secondary to +test and Stest biases.

How Information is Integrated: Reexamining the Frequency-Probability (F-P) Distinction

By focusing on whether ΔP outperforms various linear heuristics, researchers have assessed whether people tend to adopt information integration strategies that combine either probabilities or frequencies. The conclusiveness of this research, however, has been limited because the rules examined often have varied in terms of other feature distinctions. Thus, it has been difficult to determine whether the better accuracy of, for example, ΔP over $A–B$ is due to the conditional-probability structure of the former rule or some other feature that it possesses (e.g., ΔP uses four frequencies whereas $A–B$ uses two frequencies). Stated differently, both the frequency and the conditional-probability rules that have been evaluated represent a rather piecemeal rule subset. Even from a rule-analytic perspective, plausible models have gone untested. For example, given the evidence in favor of a +test bias, one might expect that $P(O_p|I_p) - P(I_a|O_p)$ would outperform ΔP in terms of its viability. That is, the contrast of two +tests might be expected to outperform the contrast of a +Itest and a –Itest. From a feature-analytic perspective, a broader set of rule units is required in order to test the feature distinctions under investigation.

Another problem has been that, until recently, the viabilities of joint probabilities have not been investigated. Moreover, the two studies (Anderson & Sheu, 1995; Kao & Wasserman, 1993) that have done so, have used rules with different weighting structures, and in both cases the weights were determined empirically by model fitting group-averaged ratings. In contrast, the approach taken here is to construct basic rules that possess specific conjunctions of features. Table 1 shows the notation and the computations for the 12 rules representing combinations of P-N, I-O, and F-P distinctions (the specification of cumulative strategies was provided earlier in the section on the C-C distinction). Frequency rules (Ftests) and joint-probability rules (Pjtests) are structured so that a positive value indicates a facilitative relation and a negative value indicates an inhibitory relation; conditional-probability rules (Pctests) are structured so that values greater than .5 represent facilitative relations and values less than .5 represent inhibitory relations. These conventions, however, do not affect rule viability [e.g., $A - B$ and $B - A$ are equally viable; the same goes for $A/(A + B)$ and $B/(A + B)$].

The F-P Distinction as an Issue of Standardization over Different Focal Sets

Theoretically, the distinction between Ftests, Pjtests, and Pctests has not been made clear. I interpret this distinction in terms of the issue of standardization over different focal sets. The predicted values of standardized tests fall within a range defined by two constants. Both Pjtests and Pctests are standardized tests. That is, Pjtest values range between -1 and $+1$ whereas Pctests range between 0 and 1 (the fact that the constants for Pjtests and Pctests differ is trivial and can be equated by a linear transformation). In contrast, Ftests are unstandardized. Theoretically, their values range between $-n$ and $+n$, where n is the size of a *focal* set. Whereas N represents the overall set size (i.e., the sum of the four cell frequencies), n may represent a subset of N. For example, if $A = 5$, $B = 2$, $C = 1$, $D = 2$, then $N = 10$, but if a judge focuses only on information from Cells A and B, then $n = 7$.

Pjtests and Pctests represent two different types of standardized Ftests. In the case of Pjtests, this is transparent: A linear contrast between two frequencies is standardized by N. That is, a Pjtest is simply an Ftest divided by n, where $n = N$. In the case of Pctests, the corresponding Ftest is standardized by the frequency of cases within the row or the column over which a contrast is computed. I refer to this restricted set of cases as an *event* set. Thus, for Pctests, $n \leq N$. This relation between Pctests and Ftests may at first be unclear because Pctests, in fact, are linear transformations of perhaps more intuitive representations of event standardization. For example, consider F(+I): The focal set for this contrast is the I_p event set. The intuitive event-standardized representation of F(+I) is $(A - B)/(A + B)$, but

$$(A - B)/(A + B) = A/(A + B) - B/(A + B) = A/(A + B) - [1 - A/(A + B)] = 2A/(A + B) - 1,$$

which simply is a linear transformation of Pc(+I) of the form $y = 2x - 1$, and which therefore will have the same viability as Pc(+I) in any focal set.

Proposition 5: Judges Seek Comparability Across Inferential Contexts

There is reason to believe that judges adopt standardized test strategies. If they relied on Ftests it would be difficult to interpret computational values because the scale on which those values lie would vary across focal sets of different sizes. For example, consider a judge who uses

F(+I) in four different stimulus sets (S1, S2, S3, S4): in S1, $A = 2$, $B = C = D = 0$; in S2, $A = 11$, $B = 9$, $C = D = 0$; in S3, $A = 2$, $B = 0$, $C = D = 10$, and in S4, $A = 11$, $B = 9$, $C = D = 10$. For each set, F(+I) = 2. However, in S1, this result represents all the cases in the event and overall sets, whereas in S2 this result represents a small fraction of cases in these two sets. In S3, the same result represents all the cases in the event set but only a small fraction of cases in the overall set. Finally, in S4, that difference represents a small fraction of the event set and an even smaller fraction of the overall set. In contrast to F(+I), both Pc(+I) and Pj(+I) yield values whose magnitudes can be interpreted unambiguously because the scale remains constant despite changes in focal set size across different situations (i.e., the value changes rather than the scale). In psychological terms, standardization permits *comparability* across focal sets of differing size.

It is unclear, however, whether judges would consistently opt for either Pctests or Pjtests. Pctests are intuitively appealing because the focal set consists solely of those cases that are directly relevant to a given contrast. Thus, the event set size might be viewed as the most relevant basis for establishing comparability. However, in situations where a focal event set over which a contrast is computed is itself a subset of a salient overall set, the larger set size N may be selected as the basis for establishing comparability. The reason is that, unlike Pctests, Pjtests are sensitive to the marginal probability of the focal event. All other things being equal, as this probability decreases, Pjtest values also decrease [e.g., compare Pj(+I) for S1 and S3 in the previous example: the values for S1 and S2 = 1 and .09, respectively]. That is, in any given situation, the computed value of a Pjtest is determined solely by the marginal probability of a focal event set and the ratio of the contrast difference value to the marginal frequency of the focal event set (viz., the Pctest value). I examine the viabilities of Ftests, Pctests, and Pjtests in Experiment 3 and in a reanalysis of data from five past experiments.

Summary

In contrast to the rule-analytic framework that characterizes the past 30 years of research on information integration in causal inference, I have proposed an alternative feature-analytic framework that asks the question, "What are the important features that describe information integration processes?" This new framework has some advantages. First, by describing rules as conjunctions of features, it offers a theoretical account of why inconsistencies in rule rankings may be expected to occur. Second, this framework lends itself methodologically, through feature analysis, to an extension of standard correlational techniques that refocuses attention on building models by using principles rather than simply viewing model viability as a function of predictive utility. Third, I use this framework to develop a feature-based account of causal inference that focuses on (a) the type of information that is integrated and (b) how, at a structural level, that information is combined. In addressing the first of these two foci, I show how the stable findings concerning how people weigh information in causal inference tasks can be explained in terms of a small set of feature preferences—namely, a primary tendency toward +testing and a secondary tendency toward Stesting (a somewhat weaker tendency toward Itesting is also predicted). This account not only provides a theoretical basis for interpreting one of the most (if not the most) invariant set of findings in this research domain, it also serves a theoretically integrative function, suggesting both commonalties and differences between rule-discovery and causal-inference processes. In addressing the issue of how information is integrated, I recast the F-P distinction in broader terms that include Pjtests and reconceptualize this distinction as one that concerns the issue of focal set standardization (in computational terms) or comparability (in psychological terms). In the next section, both the methodology and the theoretical predictions developed previously are put to empirical tests.

EMPIRICAL TESTS

Overview

In three experiments, subjects were asked to characterize the effect of an antecedent event on a consequent event based on their assessments of numerically summarized contingency data. Following the causal-rating task, subjects provided self reports of their strategy use. The main goal of the research was to test the viability of the present theoretical account using feature-analytic techniques. Specifically, I tested the four propositions made earlier concerning the type of information that is integrated. In addition, in Experiment 3 and in a reanalysis of data from Anderson and Sheu (1995, Experiment 1), Kao and Wasserman (1993, Experiment 2), Levin et al.

(1993, Experiment 1), and Wasserman et al. (1990, Experiments 2 and 3), I examined the relative viabilities of Ftests, Pjtests, and Pctests.

As noted earlier, in order to gain insight into the reasons why subjects respond as they do, it is important to examine their definitions of key concepts such as causality or preventability. Mandel and Lehman (in press) suggest that people interpret the term *cause* more specifically to mean *facilitative* cause and refer to *inhibitory* causality as *preventability*. In Experiment 1, subjects were asked to define the terms *cause* and *preventor*, thus providing a basis for testing this hypothesis. These data are also relevant to testing the prediction that judges are more concerned with sufficiency than with necessity. Presumably, these concerns would be reflected in subjects' conceptions of causation.

Another issue examined is the effect of descriptive content on causal judgments. Most past studies have used scenarios involving events for which most people would readily be able to access relevant world knowledge (e.g., the relation between the use of fertilizer and plant blooming, the relation between drug use and a side effect of the drug). Given that prior beliefs about the relation between events can bias judgments of new data (Alloy & Tabachnik, 1984; Jennings et al., 1982; Trolier & Hamilton, 1986), it is important to examine how people assess the relation between abstract events. Therefore, abstract events were used in Experiments 1 and 2; a more thorough test of the effect of content was undertaken in Experiment 3 in which two content conditions and an abstract condition were compared (each condition used identical data configurations).

Finally, I examined the effect of information presentation on information integration. The present experiments used a numeric summary format for presenting data. This format was adopted because, as noted earlier, it minimizes the impact of cognitive demands that are not part of the information integration process *per se* (e.g., memory and judgment demands required in recalling and mentally tabulating instances when data are presented serially rather than simultaneously). To test the generalizability of the obtained findings, however, I reanalyzed data from Anderson and Sheu (1995, Experiment 1) and Kao and Wasserman (1993, Experiment 2, discrete-trial condition) which were collected using a discrete-trial format.

Method

Subjects

The subjects in Experiments 1-3 were University of British Columbia undergraduate students who received course credit for participating. Sample sizes were 52 (40 female, 12 male; mean age = 21.4 years), 51 (41 female, 10 male; mean age = 21.3 years), and 120 (85 female, 35 male; mean age = 20.8 years) subjects for Experiments 1-3, respectively.

Procedure

In Experiments 1 and 2, subjects first were told that the experimenter is "interested in learning about how you think about relationships between abstract events. Your task in this study is to assess the effect of 'Event A' on 'Event B'." Subjects were told to consider the information presented to them carefully but not to spend too much time trying to analyze the information. Contingency information was presented as in Figure 2, after which subjects were asked, "How would you characterize the effect of Event A on Event B?" Subjects responded on a 7-point Likert scale ranging from −3 (Event A is a strong preventor of Event B) through 0 (Event A neither prevents nor causes Event B) to +3 (Event A is a strong cause of Event B), with the absolute values of 1 and 2 anchored at "weak" and "moderate," respectively.

In Experiment 3, subjects were randomly assigned to one of six conditions in a 2 (Information Order: A-D vs. D-A) × 3 (Event Type: abstract vs. content-natural vs. content-social) factorial design. In the A-D conditions, subjects received cell frequency information in the order *A*, *B*, *C*, *D* as in Experiments 1 and 2 (see Figure 2). In the D-A conditions, the order was reversed. Subjects in the *abstract* conditions followed the same procedure as in Experiments 1 and 2. In the *content-natural* conditions, subjects read a modified version of Kao and Wasserman's (1993) instructional set (see Appendix A) in which they were asked to evaluate, in succession, the effects of 36 experimental fertilizers on plant blooming. Finally, subjects in the *content-social* conditions were asked to evaluate, in succession, the effects of 36 different personality traits on "willingness to initiate a conversation with a stranger" (see Appendix A).

Subjects responded using the scale employed in the previous experiments, except that appropriate event labels were substituted in the content conditions.

In all three experiments, subjects completed a post-experimental survey. Subjects 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. If they indicated having used a strategy, they were asked to describe it. They were then asked whether or not they "mapped" the abstract events used in the experiment onto more meaningful real-world events in order to answer the questions being asked (this question was not asked in the content conditions in Experiment 3). If they responded affirmatively, they were asked to describe these events. In Experiment 1, subjects were also asked to define the terms *preventor* and *cause*. Following the post-experimental survey, subjects were debriefed and thanked for their participation.

Stimulus Sets

The stimulus sets used in Experiments 1 to 3 are shown in Tables 2 to 4, respectively. In Experiment 1, each of the 14 stimuli yielded a particular combinatorial pattern of conditional probabilities that were either high (H) in value (viz., $.9459 \leq P \leq .9998$) or low (L) in value (viz., $.0002 \leq P \leq .0541$).[5] For each of these stimuli, $N = 100{,}000$. In Experiment 2, the 35 stimuli represent all combinations of 4 cases within the four cells; namely, combinations of one 4 and three 0s (4C1 = 4), combinations of one 3 and one 1 (4C1 × 3C1 = 12), combinations of two 2s (4C2 = 6), combinations of one 2 and two 1s (4C1 × 3C1 = 12), and the combination of four 1s (4C4 = 1). In Experiment 3, the 36 stimuli consist of 12 frequency distributions each of which is crossed with three different set sizes (*N*s = 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. In each experiment, across stimuli, the four cells are matched in mean frequency. Therefore, differential impact of the cell frequencies on causal ratings would be due to subjects' differential weighing of them rather than their objective values. In Experiments 1 and 2, the order of stimuli was counterbalanced using a Latin-square design. In Experiment 3, the order of stimuli was random.

Results and Discussion

Sample Characteristics

In Experiments 1-3, respectively, 77%, 86%, and 67% of subjects had never taken statistics; 86%, 78%, and 67% of subjects reported having consciously used a strategy in arriving at their ratings; and 77%, 66%, and 85% (in the *abstract* condition) of subjects claimed *not* to have mapped the abstract events onto real-world events. Mean causal ratings (both across and within stimuli) did not differ reliably as a function of any of these three variables, or of gender, in any of the three experiments (smallest $p = .12$). Thus, analyses in Experiments 1-3 are collapsed across these groupings.

Preliminary Statistics

The mean causal ratings for Experiments 1-3 are shown in Tables 2-4, respectively (the ratings in these experiments were divided by 3 thus providing a possible range of −1 to +1). The grand means for these experiments are −.02 (*SD* = .49), .04 (*SD* = .42), and −.01 (*SD* = .42), respectively. The fact that none of these values differs reliably from the scale midpoint of 0 indicates that, in these experiments, subjects were not biased toward perceiving facilitative (positive) relations more readily than inhibitory (negative) relations—a tendency that Kareev (1995) referred to as a positive correlation bias. Rather, subjects' overall ratings accurately reflect the fact that, in each experiment, the mean frequencies of CFAC and CINH cases were equal.

The Type of Information that is Integrated

Subjective Cell Importance

The predictions made earlier concerning the type of information that judges integrate follow from a robust set of findings concerning the subjective importance or weight that judges attach to data from each of the four contingency cells. The SCI findings indicate that people weigh $A > B > C > D$. Thus, I begin by examining the replicability of this pattern within the current data sets. To do so, a viability estimate (viz., r') for each of these four frequency variables was calculated for each subject.

Table 5 shows the mean viability estimates taken across subjects for the four frequency variables in each experiment. Note two aspects of the data: First, in each experiment, the valence of the mean viabilities is positive for *A* and *D* (viz., CFAC cases) and is negative for *B* and *C* (viz., CINH cases). Consistent with results of Lipe's (1990) meta-analysis, the present finding indicates that, on average, the implications of the four cells that subjects drew are consistent with logical accounts. Second, in each experiment, the absolute magnitude of the mean viabilities was greatest for *A* followed by *B*, *C*, and *D*, respectively. Replicating past SCI research, this ranking indicates that subjects weighed $A > B > C > D$. To assess this weighting inequality in a more quantitative manner, I first reversed the sign of the viability estimates for *B* and *C* (i.e., by multiplying each by −1) and then conducted polynomial contrasts on the viabilities for the four frequencies (in each experiment, the omnibus *F* tests were reliable, *p*s < .001). As a visual inspection of the viabilities in Table 5 indicates, in each experiment the linear contrast (i.e., moving from the mean viability of *A* to the mean viability of *D*) was strongest. Parameter estimates for the linear trend were −.62 (99% *CI* = −.76, −.48), −.40 (99% *CI* = −.49, −.31), and −.38 (99% *CI* = −.42, −.33) for Experiments 1-3, respectively.

Subjective representation of cell frequencies. Research in this area generally has assumed that subjects' representations of numbers reflect the objective values presented to them. However, using other paradigms (e.g., random number generation) 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 function (Banks & Hill, 1974; Moyer & Landauer, 1967) or by a decelerated power function with an exponent of about 2/3 (Banks & Hill, 1974). If subjects do transform the frequencies presented to them in such a manner, then the transformed frequencies should correlate higher than the actual frequencies with their causal ratings. I examined log and power (exponent of 2/3) transformations of the four cell frequencies. In each experiment, the actual frequencies correlated either the same or better than either of the transformations. Therefore, in subsequent analyses, the actual frequencies are used in calculating rule predictions.

Effects of order on SCI. It is important to rule out a plausible alternative explanation for the previous SCI findings. In Experiments 1 and 2, information was presented in the order of Cells A, B, C, and D, respectively—exactly the same order of the cell-weighting inequality. Therefore, that result might be due to an order (primacy) effect. In Experiment 3, half the subjects received information in the same order (A-D) and the other half received information in the reverse order (D-A). If the previous findings were due to a primacy effect, we should see, in the latter case, an attenuation of the viabilities of A and B and an increase in the viabilities of C and D.

Wasserman et al. (1990) used a numeric summary format for presenting contingency data to subjects and found that order (A-D vs. D-A) had no reliable effect on their causal ratings. Similarly, as Figure 3 shows, a primacy-effect account was not supported in Experiment 3. Regardless of order, subjects weighed $\{A > B\} > \{C, D\}$. Order did interact reliably with the within-subjects Cell Frequency factor [$F(3, 342) = 3.89, p = .009$]; however, this effect was primarily due to a decrease in the viability of C, and a slighter increase in the viability of D (perhaps due to a weak primacy effect), in the D-A condition. Though reliable, this interaction is of little theoretical significance and accounts for little variance in viabilities ($\eta^2 = .03$) relative to the effect of Cell Frequency [$F(3, 342) = 201.95, p < .001, \eta^2 = .64$]. Thus, subsequent analyses are collapsed across Order.

Effects of event type on SCI. Another goal of Experiment 3 was to examine the generalizability of SCI and other findings across event types. Thus, in addition to the abstract events used in Experiments 1 and 2, two content conditions were included. One of these dealt with events in the natural world (viz., fertilizers and plant blooming); the other dealt with events in the social world (viz., personality traits and willingness to start a conversation with a stranger). Figure 4 shows the breakdown in mean viabilities by event and cell frequency. As can be seen, Event and Cell Frequency factors interact [$F(6, 342) = 5.06, p < .001, \eta^2 = .08$]. Two aspects of this interaction are noteworthy: First, subjects weigh A greater in the content conditions, especially when the events are social in nature. Second, when the events are social in nature, subjects relied exclusively on data from Cells A and B (with a greater emphasis on Cell A). Both of these findings cohere with the present account. They suggest that +test biases are somewhat stronger when judges process content-based events that have some real-world meaning. And, in the case of social events, +Itesting is especially dominant. These findings indicate that differences between event domains is more a matter of the degree than of the type of effect that will obtain. Thus, subsequent analyses in this section are collapsed across event type.

Individual differences in SCI. Although the preceding analyses involve correlational data obtained at the subject level, the account provided is nomothetic (i.e., how does the group behave?). To provide a sense of individual differences in SCI, I ranked the viabilities of the four cell frequencies for each subject in each experiment from 1 to 4 (in some cases there were ties). Table 6 shows the percentage of cell-frequency viabilities at each rank. The most frequent ranking for *A*-viabilities was 1 followed by 2, for *B*-viabilities it was 2 followed by 1, for *C*-viabilities it was 3 followed by 4 (vice versa in Experiment 3), and for *D*-viabilities it was 4 followed by 3. Conversely, the frequency viability ranked first most frequently was *A* followed by *B*, for Rank 2 it was *B* followed by *A*, for Rank 3 it was *C* followed by *D* (vice versa in Experiment 3), and for Rank 4 it was *D* followed by *C*. These findings reveal not only that people, *on average*, weigh $A > B > C > D$, but that *most* people weigh the data in a manner consistent with this inequality.

Testing Proposition 1

Proposition 1 posits that judges consider evidential tradeoffs and, thus, should favor contrastive tests over cumulative ones. To test this hypothesis a number of C-C test pairs were pitted against each other. Specific comparisons include the viabilities of (a) F(+I) versus $f(I_p)$, (b) F(–I) versus $f(I_a)$, (c) F(+O) versus $f(O_p)$, (d) F(–O) versus $f(O_a)$, and (e) *P*(CFAC) (recall the point made in Footnote 2 that this test is contrastive) versus the mean viability of *f*(CFAC) and *f*(CINH). Because *N* was held constant in Experiments 1 and 2, $f(I_p)$ and $f(I_a)$, $f(O_p)$ and $f(O_a)$, *and* *P*(CFAC), *f*(CFAC) and *f*(CINH) are equally viable. Thus, in these two experiments the number of contrasts is reduced to two: (a) the mean viability of F(+I) and F(–I) versus $f(I_p)$ and (b) the mean viability of F(+O) and F(–O) versus $f(O_p)$. One-way (C-C Distinction) within-subjects MANOVAs revealed that contrastive tests were more viable than cumulative ones in Experiment 1[$F(2, 50) = 31.40, p < .001, \eta^2 = .56$], Experiment 2 [$F(2, 49) = 81.60, p < .001, \eta^2 = .77$], and Experiment 3 [$F(5, 114) = 333.73, p < .001, \eta^2 = .96$]. Moreover, in none of the nine univariate analyses conducted across the three experiments was a cumulative test more viable than the corresponding contrastive test. Taken together, these findings support Proposition 1.

Testing Propositions 2-4

Proposition 2 indicates that judges are strongly biased toward +testing, suggesting that a main effect for the P-N distinction would obtain. Proposition 3 indicates that, owing to a bias toward Stesting, judges favor Itesting under +test conditions and Otests under –test conditions. That is, Proposition 3 suggests a P-N × I-O interaction. Proposition 4 indicates that judges may display a bias toward Itesting, but that it is a secondary effect that is importantly qualified by the aforementioned interaction.

Feature analysis. To test these propositions, I examined the mean viabilities of the four Ftests shown in Table 1. Table 7 shows the mean viabilities for these tests in each of the three experiments. These estimates were analyzed using three 2 (P-N) × 2 (I-O) repeated-measures ANOVAs. Supporting Proposition 2, reliable main effects for P-N indicating that +tests outperformed –tests were obtained in Experiment 1 [$F(1, 51) = 121.86, p < .001, \eta^2 = .71$], Experiment 2 [$F(1, 50) = 144.94, p < .001, \eta^2 = .74$], and Experiment 3 [$F(1, 119) = 382.22, p < .001, \eta^2 = .76$].

Supporting Proposition 3, reliable P-N × I-O interactions were obtained in Experiment 1 [$F(1, 51) = 35.38, p < .001, \eta^2 = .41$], Experiment 2 [$F(1, 50) = 23.01, p < .001, \eta^2 = .32$], and Experiment 3 [$F(1, 119) = 141.26, p < .001, \eta^2 = .54$]. As Table 7 shows, the anticipated crossover interaction emerged: Itesting was more viable than Otesting given +test constraints and the reverse was true given –test constraints. Notably, the P-N main effect was not qualified by this interaction: In each experiment, +Itests outperformed –Itests and +Otests outperformed –Otests.

Finally, supporting Proposition 4, reliable main effects for I-O indicating that Itests outperformed Otests were obtained in Experiment 1 [$F(1, 51) = 19.98, p < .001, \eta^2 = .28$], Experiment 2 [$F(1, 50) = 18.68, p < .001, \eta^2 = .27$], and Experiment 3 [$F(1, 119) = 68.96, p < .001, \eta^2 = .37$]. The data also support the notion that +test biases outweigh Stest biases which in turn outweigh Itest biases. Namely, across Experiments 1-3, the weighted mean magnitude of effect estimates (η^2) are greatest for the P-N main effect (.74), followed by the P-N × I-O interaction (.46) and finally the I-O main effect (.33). Nevertheless, each of these predicted effects accounts for substantial variance in rule viability.

To test the generalizability of these findings, I calculated the viabilities of F(+I), F(+O), F(–I), and F(–O) from the mean causal ratings reported by Anderson and Sheu (1995, Experiment 1) and Kao and Wasserman (1993, Experiment 2, discrete-trial condition). As noted earlier, unlike the present experiments which used a numeric summary format for presenting contingency data to subjects, these past experiments used a discrete-trial format, which likely challenges subjects with additional cognitive demands. Table 8 shows the viabilities for these four tests. Although comparable ANOVA techniques cannot be performed (given that the mean causal ratings represent only one case), as can be seen in the table, in both experiments the pattern clearly replicates that obtained in the present experiments. Namely, supporting Proposition 2, +tests outperform –tests. Supporting Proposition 3, Itesting outperforms Otesting given +test conditions, but Otesting outperforms Itesting given –test conditions. And, supporting Proposition 4, Itesting outperforms Otesting. Taken together, then, the present findings along with the reanalysis of published data by others strongly support Propositions 2-4.

Concept Analysis. The notion that judges favor Stesting over Ntesting is supported by the P-N × I-O interactions observed in the present experiments. To further test the plausibility of this notion, subjects' concepts of *cause* and *preventor* were examined in Experiment 1. If subjects favor Stesting, then they should be more likely to define these constructs in terms of a sufficiency criterion than a necessity criterion. For instance, subjects should be more likely to state something like "when the cause is present the effect occurs" (i.e., the cause is sufficient to produce the effect) rather than something like "if the cause is not present the effect will not occur" (i.e., the cause is necessary to produce the effect). Given that people tend to focus on positive information (a notion that is directly supported in the present experiments), sufficient facilitative causation should take the form $I_p \rightarrow O_p$, whereas necessary facilitative causation should take the form $I_a \rightarrow O_a$. Sufficient inhibitory causation should take the form $I_p \rightarrow O_a$, whereas necessary inhibitory causation should take the form $I_a \rightarrow O_p$.

Consistent with the idea that people's understanding of causality (in the broader sense) is generally I_p-constrained and consistent with only a sufficiency criterion, 71% of subjects defined *cause* in terms of $I_p \rightarrow O_p$ and 76% of subjects defined *preventor* in terms of $I_p \rightarrow O_a$. However, 22% and 10% of subjects provided definitions of *cause* and *preventor*, respectively, that clearly were *not* I_p-constrained. For example, one subject defined preventor as follows: "If Event A is present, Event B will be absent. If Event A is absent, Event B will be present." The same subject defined cause as follows: "If Event A is present, Event B is present. If Event A is absent, Event B will also be absent." In general, however, the analysis of subjects' definitions adds further support to the notion that judges favor Stesting over Ntesting.

Also note that the vast majority of subjects expressed a concept of preventor consistent with inhibitory causation (88%) and a concept of cause that more specifically referred to facilitative causality (90%). This finding supports Mandel and Lehman's (in press) hypothesis that lay concepts of causes refer primarily to facilitative causation. In fact, only three subjects (6%) expressed more general notions of cause that included both facilitative and inhibitory connotations (the remaining 4% were uncodable). This finding suggests that, in general, subjects understood the intended distinction between cause and preventor.

A Nonintegrative Reference Point for Assessing Rule Viability

The preceding feature analyses indicate what types of rules might represent people's information integration strategies better than other types. For instance, +tests are more viable than −tests, indicating that judges favor +test strategies. In these analyses, a group of rules sharing a feature is contrasted with another group sharing the opposing feature. Therefore, one group serves as a reference point for evaluating the other group. One might ask, however, whether any of a set of rules is particularly viable. To do so requires an alternative reference point for comparison. One meaningful reference point is the viability of the most highly correlated cell frequency in a given situation. The underlying logic is that if a focal rule outperforms other rules but fails to outperform the viability of a single cell-frequency variable, then clearly it still has failed the test. At the least, it would be shown to lack parsimony. As in each of the present experiments, the most highly correlated frequency is usually A.

In Experiments 1-3, the mean viabilities of *A* are .82, .70, and .50, respectively. Examining Table 7, we see that, in each experiment, only the +Itest has a greater mean viability. The increment in mean viability is reliable in Experiment 1 [paired *t*(51) = 3.27, *p* = .001], Experiment 2 [paired *t*(50) = 5.71, *p* < .001], and Experiment 3 [paired *t*(119) = 17.61, *p* < .001]. This analysis indicates that, of the four contrastive Ftests, only F(+I) surpassed the viability of using *A* alone as a "nonintegrative model" of subjects' causal ratings. The only other frequency rule that surpassed the viability of *A* was *A+D* (mean *r′* = .89) in Experiment 2 [paired t(50) = 4.10, *p* < .001]; however, the viability of *A+D* did not differ from that of *A* in Experiment 1 (*t* < 1) and was reliably *weaker* (mean *r′* = .34) than that of *A* in Experiment 3 [paired t(119) = 10.19, *p* < .001]. Therefore, *A+D* lacks consistency in outperforming *A*, whereas F(+I) does not.

How Information is Integrated

Testing Proposition 5

Proposition 5 states that standardized tests (viz., Pctests and Pjtests) should outperform unstandardized tests because the former permit judgment comparability across different data contexts. In Experiment 3, in which a well-defined overall set is salient (owing to the fact that data for the four cells are numerically summarized), Pjtests may be expected to outperform Pctests as well. To test this notion, the mean viabilities of Ftests, Pctests, and Pjtests were compared across four measures (viz., +Itest, +Otest, −Itest, −Otest), while taking into account the potential effect of event type as well. A 3 (Event) × 3 (F-P) mixed MANOVA on these viabilities revealed a reliable F-P main effect, *F*(8, 110) = 44.42, *p* < .001, η^2 = .76 (as well, all four univariate *F* tests were reliable, *p*s < .001). This effect was clarified by a reliable Event × F-P interaction, *F*(16, 220) = 4.34, *p* < .001, η^2 = .24. However, as can be seen in Table 9 (top section) which shows the mean viabilities (collapsed across the four measures) for Ftests, Pctests, and Pjtests, the viability patterns for the three Event subgroups all led to the same basic finding that Pjtests were most viable in this data context. This finding supports Proposition 5.

To examine this issue further, I reanalyzed data from Anderson and Sheu (1995, Experiment 1), Kao and Wasserman (1993, Experiment 2), Levin et al. (1993, Experiment 1), and

Wasserman et al. (1990, Experiments 2 and 3).[6] Like the present experiments, the reanalyzed experiments by Kao and Wasserman (summary condition), Levin et al., and Wasserman et al. used numeric summary formats for presenting data. However, as noted earlier, the reanalyzed experiments by Anderson and Sheu and Kao and Wasserman (discrete-trial condition) used a discrete-trial format. The overall set size may be especially salient when data are presented in a numeric summary format. Earlier I hypothesized that Pjtests may outperform Pctests when an overall set is salient, but that Pctests may outperform Pjtests when the event set but not the overall set is salient. If this is correct, the following pattern of viabilities should be observed: Pctests > Pjtests > Ftests under discrete-trial procedures and Pjtests > Pctests > Ftests under summary procedures. I calculated viabilities for each of the 12 rules shown in Table 1 based on the mean causal ratings reported by Anderson and Sheu. Table 9 shows the viabilities for Ftests, Pctests, and Pjtests, collapsing across the P-N and I-O distinctions.

As can be seen in the table, the data cannot be fully accounted for in terms of the preceding hypothesis. Although Anderson and Sheu's data fit the prediction for discrete-trial procedures (Pctests > Pjtests > Ftests), Kao and Wasserman's trial-condition data led to the exact opposite pattern. And, whereas the data from my Experiment 3 fit the prediction for summary procedures (Pjtests > Pctests or Ftests), the data from other experiments employing numeric summary procedures did not. Namely, the Levin et al. data indicate that Pctests ≅ Pjtests > Ftests; the Wasserman et al. data indicate that Pctests > Pjtests = Ftests; and Kao and Wasserman's (summary condition) data indicate that Ftests > Pjtests > Pctests. With the exception of Kao and Wasserman's experiment, the remaining experiments support Proposition 5. Nevertheless, compared with the impeccable consistency of the results supporting the other propositions, these findings raise the question of whether the F-P distinction is useful for drawing out invariances in judges' information integration strategies. It may be the case that biases on this distinction are locally determined by a set of factors including the nature of the data themselves and the manner in which the data are received.

A Tradeoff Between Sample Size and Comparability Considerations

Proposition 5 indicates that judges would favor Pctests or Pjtests over Ftests because only the former allow for comparability across data contexts. Achieving comparability, however, is at odds with the goal of being sensitive to changing sample sizes. This is one of the reasons why Anderson and Sheu (1995) have argued that ΔP is *not* normative. More generally, the values of Pctests and Pjtests do not change as a function of *N* when the distribution of cell frequencies is held constant. Clearly, it is reasonable that judges should be more confident in their inferences when those inferences are based on a larger, more reliable set of data. Therefore, neither Pctests nor Pjtests are attuned to the law of large numbers.

More important is the question of whether judges are sample-size insensitive. Anderson and Sheu (1995) found only a weak effect of sample size on the extremity of subjects' causal ratings for problem sets in which either *A*, *B*, *C*, or *D* was three times as large as the other three values. They concluded that, as has been found in other judgment domains (e.g., Tversky & Kahneman, 1974), subjects are sample-size insensitive. As noted earlier, Anderson and Sheu used a discrete-trial format which brings into play other cognitive processes that are not part of information integration process *per se*. In Experiment 3, these additional demands are removed. Therefore, if subjects display sample-size neglect in that context, we can be more certain that it is not due to processing constraints imposed by other cognitive demands. As noted earlier, the stimuli in Experiment 3 consist of 12 distinct frequency distributions that are crossed with 3 different overall set sizes (see Table 4). In each quarter of the 12 distributions, a distinct cell frequency (Large Cell) is either 2.5 or 5 times larger than the remaining three frequencies. If subjects are sensitive to the overall set size, then as *N* increases, their ratings should become more extreme (a) in the positive direction when *A* or *D* is the largest frequency (A-large and D-large) and (b) in the negative direction when *B* or *C* is the largest frequency (B-large and C-large).

Figure 5 shows the pattern of mean causal ratings grouped by Large Cell as a function of set size (for more detail, see Table 4 for exact mean values for each stimulus as a function of set size). There was a reliable trend for causal ratings to become more extreme in the expected direction as *N* increased for A-large [$F(6, 114) = 11.95$, $p < .001$, $\eta^2 = .386$] and for B-large [$F(6,$

114) = 12.56, $p < .001$, $\eta^2 = .398$]. However, the effects of *N* on mean ratings of C-large and of D-large were not reliable [*F*s(6, 114) = 1.82, $p = .11$]. These findings indicate that subjects are somewhat sensitive to changes in *N* at least in situations where *A* or *B* represent the bulk of cases. In both of these situations, set size explained over a third of the variance in causal ratings within Large-Cell stimulus groupings. Nevertheless, whereas *N* increased by a factor of 4, mean causal ratings for A-large and B-large increased in extremity by a factor of roughly 1.5.

The findings suggest that, although subjects do take overall set size into account in determining the extremity of their judgments, they attach more importance to maintaining judgment comparability across data contexts. Given that subjects were asked to assess the *effect* of the antecedent on the consequent rather than assess their *confidence* in their judgment, the manner in which they appear to have resolved the tradeoff in considerations may be viewed as rational. Had subjects been asked to rate their confidence in their judgments (a focus of future research), one might find subjects' confidence judgments to be considerably more sensitive to changes in *N*.

Self Reports of Strategy Use

Differences in self reports emerged between the experiments. Although a reliably greater proportion of subjects in Experiment 1 (86%) than in Experiment 3 (67%) claimed to have used a strategy [$\chi^2(1, N = 172) = 7.21$, $p = .007$], the accounts provided in the first two experiments were much more vague than those provided in Experiment 3. As well, among subjects who did provide accounts that could be translated into a decision rule, the types of strategies purportedly employed differed across the experiments.

In Experiment 1, one distinct cluster of self reports emerged: Twenty-two subjects (42%) claimed to adopt one of two related strategies: (a) an *anchoring* strategy in which they focused solely on the largest cell frequency (relating the assumed implication of that cell to the scale) or (b) an *anchor-and-adjust* strategy (Einhorn & Hogarth, 1986; Tversky & Kahneman, 1974) in which they determined the causal direction by the implication of the largest cell frequency but determined the magnitude by adjusting for the value of one or more of the remaining frequencies. Not even one subject in Experiment 1 described anything that resembled the calculation of a probability–let alone the calculation and combination of two probabilities as in ΔP. However, 73% of subjects who reported using a strategy did mention using frequency information.

No dominant cluster of self-reported strategy emerged in Experiment 2. Of the few subjects that did mention codable strategies, two claimed to rely strictly on *A*, two described a confirming-cases strategy, two described an anchor-and-adjust strategy in which the largest cell frequency served as the anchor, and two claimed to rely solely on the largest frequency. One subject described computing Pc(+I), and two mentioned considering the cell frequency out of four, indicating that they were calculating joint probabilities.

In Experiment 3, two primary clusters of self-reported strategies were identified. Consistent with the reports in Experiment 1, 18 subjects (15%) claimed to use either an anchoring or an anchor-and-adjust strategy. However, in contrast to the apparent lack of use of probabilities in Experiment 1, 21 subjects (18%) in Experiment 3 described strategies that closely approximated ΔP. For instance, one subject wrote, "I tried to find the percentage of flowers that bloomed using the fertilizer versus the percentage of flowers that bloomed without the fertilizer."

As well, 15 subjects (13%) specifically mentioned the need to take *N* into account in arriving at a judgment. For some subjects, this concern was based on comparability considerations. For example, one subject stated, "Because the group numbers [i.e., *N*] varied, when comparing numbers, I used ratios." Another subject actually criticized the experimenter for not reporting the data in terms of proportions of *N*. For other subjects, however, concerns centered on reliability issues. For example, one subject wrote, "If only 10 people were surveyed, I would never answer stronger than *weak*. I never answered *strong* because I don't think 40 people surveyed is enough to measure the effects of a personality trait." Another wrote, "I later realized that I shouldn't have attached as strong a cause or preventor [label] to the data where there was only 10 instances because the results were more likely due to chance."

Reanalysis of Data from Self-Reported ΔP Users

Consistent with Propositions 2-4, it was found in each experiment that, in terms of viability, +Itesting > +Otesting > –Otesting > –Itesting. An unweighted ΔP strategy, however, implies that +Itesting = –Itesting > ±Otesting. ΔP also implies that Pctesting would be favored over Pjtesting or Ftesting. To test the accuracy of the 21 self-reported ΔP users' accounts, I reanalyzed the viability data from this subgroup. The mean viabilities for the 12 rules shown in Table 1 were analyzed in a 2 (P-N) × 2 (I-O) × 3 (F-P) repeated-measures ANOVA. Even among the group of self-reported ΔP users, a +test bias was observed [mean $r' = .70$ for +tests and .48 for −tests; $F(1, 20) = 29.10, p < .001, \eta^2 = .51$]. In further contradiction of subjects' reports, no Itest bias was observed [mean $r' = .60$ for Itests and .59 for Otests; $F(1, 20) = 1.11$, ns.] and the P-N × I-O interaction replicated the predicted pattern found in the overall sample $F(1, 20) = 12.98, p = .002, \eta^2 = .39$]. Finally, Pctests and Pjtests were equally viable (mean $r' = .63$; for Ftests, mean $r' = .52$). These findings support the propositions of the current account, even though these subjects' self reports clearly suggest otherwise. Consistent with the claims of past literature (e.g., Nisbett & Wilson, 1977) cautioning researchers to be wary of the accuracy of self reports, these findings suggest that subjects do not (and perhaps are unable to) accurately report their strategies even when the reports are provided immediately following the rating task (cf. Anderson & Sheu, 1995).

Information Integration Failure

Some subjects (particularly in Experiments 1 and 2) reported that they had difficulty arriving at a judgment because, depending on what type of information they considered, different responses seemed appropriate. For example, one subject in Experiment 1 suggested correctly that *A* and *D* support facilitative causation whereas *B* and *C* support inhibitory causation. The subject, however, went on to say, "I became confused sometimes because a strong cause can also be seen as a weak preventor and therefore sometimes I couldn't decide which number to circle." Similarly, a subject in Experiment 2 wrote, "The results sometimes seem to offset each other. For instance, 'Event A occurred and Event B occurred two times *and* Event A occurred and Event B did not occur two times.' Thus, Event A is a strong cause as well as a strong preventor of Event B. Hence, I didn't know which one to pick, or should I choose neither a preventor nor a cause?" Indeed, this type of confusion would arise if the frequencies are not somehow integrated. For example, if *A* is relatively large, *B* is relatively small, and the implications of these frequencies are treated independently, one may be at a loss to decide whether Event A is a strong cause of Event B (i.e., based on the high value of *A*) or whether Event A is a weak preventor of Event B (i.e., based on the low but positive value of *B*). In contradistinction to nonnormative integration strategies or information integration biases, I refer to the tendency of some subjects to treat cell frequencies independently in terms of their causal implications as *information integration failure*. Information integration failure results in an apparently paradoxical decision state in which both of two (or more) contradictory alternatives seem plausible simultaneously.

Data-Driven Inferences versus Data-Driven Strategies

Nisbett and Ross (1980) argued that, in assessing covariation and making inferences, people tend to be overly theory-driven and are insufficiently data-driven. That is, as others (e.g., Jennings et al., 1982) have demonstrated, people tend to weigh their personal theories more heavily than available "external" data in arriving at judgments. However, people may also be data-driven in another sense, which is that the information integration strategies they adopt might vary from one problem to another. Or one might rely on a single strategy that uses different variables depending on their values in particular situations. In this sense, data-driven strategies differ from all of the rules previously discussed (e.g., ΔP, ΔD, Cell-A strategy, etc.), which use the same variables regardless of what the data look like. And, the variables that are used, in turn, are combined in a consistent manner.

The anchoring and anchor-and-adjust strategies described earlier provide examples of data-driven strategies. These strategies, in fact, are criteria for determining which variables to consider. For instance, the anchoring strategy described by many subjects provides a decision rule for determining which cell to focus on. Namely, the cell with the highest frequency determines the judged causal direction via a set of assumed implications which may or may not coincide with logical accounts.

Other independent decision rules may be invoked in order to arrive at a magnitude assessment. Another example of a data-driven strategy described by some subjects in Experiment 3 is to calculate two or more contrasts and then to use the one that yields the highest value. For instance, a judge might calculate F(+I) and F(+O): If F(+I) = 2 and F(+O) = 4, the judge would draw conclusions based on F(+O); however, had the values been reversed, the same judge would rely on F(+I). Users of this type of strategy (viz., maximal contrast selection) would be prone to overestimating causality because they would always favor tests that minimize the chance of inferring a noncausal relation. The tendency of some subjects to choose tests that maximize a contrast value may reflect an erroneous lay statistical intuition about diagnosticity. Namely, if the contrast value is greater, the test tells you more (when really all it does it tell you something different). Perhaps the resistance to publishing nonsignificant hypothesis-test results, and the apparent excitement over finding that $p < .001$, are examples of an analogous reasoning error that occurs in some scientific thinking.

Summary

The present findings support the five major propositions of the present feature-based account. As in past research, subjects in each experiment weighed $A > B > C > D$, as evidenced by the correlational structure of these variables with subjects' causal ratings. The dominance of Cells A and B was evident even when information was presented in the D-A order, and this basic SCI finding was upheld across both abstract and content-based event types. The contrast between the viabilities of frequency rules and of A (a nonintegrative reference point) revealed that only F(+I) consistently outperformed this frequency. Again, this finding coheres with the present account, which was further supported by an analysis of subjects' concepts of causation in Experiment 1 revealing a tendency to provide I_p-constrained definitions of both *cause* and *preventor*. Feature analyses of data from Experiment 3 and qualitative analyses of data from past studies generally supported the notion that judges favor standardized tests. However, the conjecture that Pjtests might outperform Pctests when N is made salient but that Pctests might outperform Pjtests when n = the event set size was not supported. Despite the fact that Ftests fared poorly by comparison in Experiment 3, set size did account for a substantial portion of the variance in mean causal ratings of stimuli grouped by the largest cell frequency. This finding may result from tradeoffs made between comparability and reliability considerations. Finally, self-report data revealed intuitive shortcomings on the part of some subjects, including information integration failure (i.e., the tendency to treat each piece of evidence independently) and a tendency to favor data-driven strategies (e.g., anchoring, maximal contrast selection).

GENERAL DISCUSSION

Consistency and Inconsistency in Information Integration

Early on I addressed the question of why rule-analytic research has not achieved its goal of finding the "best" rule that accounts for people's information integration strategies in causal inference (viz., finding "the winner"). From a feature-analytic perspective, I conceptualized a rule as a conjunction of features. I anticipated that some features clearly contribute to (or detract from) rule viability because they map onto relatively invariant information processing tendencies. Other features, however, may not be so closely tied to cognitive consistencies. The distinctions that these features comprise, therefore, would not lead to consistent results in terms of their ability to account for people's judgments. Given that a rule meshes some stable diagnostic features with some unstable diagnostic features (as well as possibly some nondiagnostic features), a basic set of "winners" might be identifiable but among that set no stable ranking will emerge. This instability owes to the fact that judges have unstable preferences in terms of some feature distinctions. In effect, then, even the most consistent winners are bound to lose sooner or later. Alternatively, if we focus on features rather than rules, relatively stable winners can be identified.

The preceding notions are supported by the present research which revealed that while C-C, P-N, and I-O feature distinctions accounted for substantial variance in rule viability in a consistent manner across experiments, the F-P distinction did not. Nor did this inconsistency stem from any obvious moderating factor. Rather it appears that, across studies, subjects were not cognitively bound to combine contingency data in a consistent manner, even though within each specific experimental context preferences for Ftesting, Pctesting, or Pjtesting did arise. One advantage of the present feature-analytic approach is the ability to isolate sources of inconsistency in rule ranking because rule viability is partitioned along the lines of theoretically meaningful feature distinctions. Past research was unable to do so because the rules being compared differed simultaneously in terms of several different distinctions.

Weighting and Combination: Isolating the Locus of Consistency in Information Integration

Interestingly, I found that consistency lies primarily in the *type* of information that subjects tend to combine rather than in *how* they go about combining it. In general, subjects relied on data sources that had contrasting implications. Specifically, they relied primarily on +Itesting, which always provides information on the sufficiency of a causal candidate. Indeed, this was the only test type that consistently outperformed *A* alone in terms of viability. What judges do not do, however, is consistently rely on either Ftesting, Pctesting, or Pjtesting. Perhaps judges know what type of data they are looking for to answer a causal question but are unsure of how best to combine it. As Anderson and Sheu (1995) suggested, in the absence of a clear strategy, judges may be influenced by the salience (or, alternatively, the simplicity) of some combination methods over others in a given context.

The distinction between the type of information that is integrated and how that information is integrated maps onto the two basic components of the information integration process: weighting and combination (Anderson, 1981). All rules specify both a weighting and a combination method, even if no weights are represented explicitly. For example, Equation 1 shows that ΔP subtracts one conditional probability from another. However, Equation 1 represents a special case of the following general conditional probability rule

$$\Delta \mathrm{Pc} = \omega_1 \mathrm{Pc}(+\mathrm{I}) + \omega_2 \mathrm{Pc}(-\mathrm{I}) + \omega_3 \mathrm{Pc}(+\mathrm{O}) + \omega_4 \mathrm{Pc}(-\mathrm{O}) - c, \qquad (2)$$

in which $\omega_1 = \omega_2 = 1$ and $\omega_3 = \omega_4 = 0$ (ω_i are probability weights) and $c = 1$ (if we are concerned only about viability but not making the possible range of values symmetric round 0, c may be omitted from the equation). Therefore, the distinction often made between weighted and unweighted rules is technically incorrect. The real distinction is between rules that are weighted a priori on theoretical grounds versus rules that are weighted post hoc based on empirical analyses of a sample of data. A related point is that rules that are treated as distinct may be viewed as the same rule with different weights assigned to the units. For example, *A–B*, the confirming-cases strategy, and ΔD are all special cases of a single, but more general, frequency rule

$$\Delta \mathrm{F} = \omega_A A - \omega_B B - \omega_C C + \omega_D D, \qquad (3)$$

in which all four weighted cell frequencies are included. For example, if $\omega_A = \omega_B = 1$, $\omega_C = \omega_D = 0$, $\Delta F = A-B$; if $\omega_A = \omega_D = 1$, $\omega_B = \omega_C = 0$ or if $\omega_A = \omega_D = 0$, $\omega_B = \omega_C = 1$, ΔF = the confirming-cases strategy; and if $\omega_i = 1$, $\Delta F = \Delta D$.

The type of information judges focus on is represented by their weighting of a set of variables. For example, the tendency to favor +Itests over other test types is formally represented by weighting some types of information (viz., *A* and *B*) more heavily than others (viz., *C* and *D*). Whereas many different weighting patterns can be ascribed to a given rule, different rules, such as Equations 2 and 3, represent a fundamental difference in how information is combined. The fact that this research (including the reanalysis of five past experiments) found consistency in terms of the type of information subjects combined, rather than how they did so, indicates that consistency resides primarily in the method of weighting information rather than in the method of combining information.

Turning Model Building on its Head

A common approach to model building is to specify precisely the combination method but to allow the weighting of the variables selected to be determined empirically by regression analyses. The findings of the present research indicate that this may not be the best approach because judges' combination methods are less consistent than their manner of weighing specific types of information. Building a model that applies across many data contexts should adhere to two simple principles: (a) account for demonstrated invariant properties of the judgment process in question and (b) avoid imposing invariant properties that, in fact, do not exist.

In light of the present findings, these two principles suggest that the standard approach to model building be turned on its head. That is, one might begin with a specification of weighting relations for different sets of variables that reflect different combination methods (e.g., frequencies, conditional probabilities, joint probabilities). Of course, this does not mean that exact relative weights need be incorporated into a model. Yet a model should integrate the finding that judges weigh $A > B > C > D$ at least by expressing this fact as a weighting inequality. From the perspective of the present account this does not only represent an acknowledgment of the robustness of this finding but of the theoretical basis for why this finding obtains.

Each of these sets of weighted variables, or *ensembles*, can also be ascribed a weight and can be combined into a single formal statement. At present, we do not know how the weighting of ensembles operates. One likely possibility, however, is that the ensembles be given unit weights so that they are either "on" (i.e., $\omega = 1$) or "off" (i.e., $\omega = 0$), and that turning one ensemble on results in the others being turned off. These constraints imply, for example, that if in a given context judges tend to use Ftests, then they won't also use Pctests or Pjtests.

For example, the following three formal statements constitute one possible model that includes three weighted ensembles corresponding to frequency, conditional-probability, and joint-probability units, respectively:

$$\mathrm{R} = b_1\{\omega_1 A - \omega_2 B - \omega_3 C + \omega_4 D\} + b_2\{\omega_1 p_1 + \omega_2 p_2 + \omega_3 p_3 + \omega_4 p_4 - c\} \tag{4}$$

$$+ b_3\{\omega_1 A - \omega_2 B - \omega_3 C + \omega_4 D\}/N + e.$$

$$\exists \mathrm{R}\ [E(|\omega_1|) > E(|\omega_2|) > E(|\omega_3|) > E(|\omega_4|)] \tag{5}$$

$$\forall b_i\ [(b = 0 \vee b = 1)\ \&\ (\Sigma b_i = 1)] \tag{6}$$

In Equation 4, R stands for an information integration response of the type examined in this research domain; p_i in the second ensemble refers to Pc(+I), Pc(+O), Pc(–O), and Pc(–I), respectively; and e stands for prediction error. In effect, Equation 4 defines a set of "players" but does not fill in the "ground rules." Statements 5 and 6 do the latter: Statement 5 expresses the weighting inequality in a probabilistic manner that applies across the three combination methods (the symbol ∃ means "there exists . . . such that . . ."; E refers to the expected value). Thus, it states that there exists R such that the expected value of the weights ω_i follow the stated inequality. Statement 6 expresses the two aforementioned constraints on ensemble weights b_i (the symbol ∀ means "for all . . . it is the case that . . . "). Thus, it states that all such weights must equal either 0 ("off") or 1 ("on") and that only one weight can be on at a time.

Taken together, Equation 4 and Statements 5 and 6 constitute a working model. Although it is more complex than most of the rules proposed and examined in past research, it is surely still an oversimplification. For example, as Anderson and Sheu (1995) observed, in the free-operant paradigm subjects' ratings were better modeled by rates than by probabilities. This suggests that additional ensembles may need to be added to Equation 4. Obviously, the rules for determining when one ensemble may be expected to be on and the others off will be an important direction for future research. Unless that goal can be achieved Equation 4 of the present model may be a more accurate representation than previous models, although it will still lack predictive value.

Confirmation, Contrast, and Nonintegration

Many social cognitive accounts of covariational assessment (e.g., Smedslund, 1963) and social inference (e.g., Fiske & Taylor, 1991; Nisbett & Ross, 1980) have posited that judges rely primarily or even exclusively on data from Cell A. From an information integration perspective, this strategy represents a nonintegrative account because only one type of information is considered. Some have suggested that the reason judges rely on Cell A is that it represents direct confirmation of their beliefs (assuming that a facilitative relation is posited). From this perspective, another strategy that judges might adopt is to focus exclusively on Cells A and D because both of these cells include CFAC cases. Although the latter strategy includes two sources of CFAC data rather than one, a tendency for judges to focus on Cell A alone may be expected because (a) it requires no integration and thus is a simpler strategy to undertake and (b) the fact that Cell D is confirmatory may not be easily apprehended by many judges. As noted earlier, the $A+D$ strategy is a cumulative one, although when N is known this strategy has contrastive implications (viz., because $N - A - D = B + C$). In essence, the account of why judges might opt for either of these strategies is based on the notion that they are prone to a confirmation bias in which disconfirmatory data is neither sought actively nor weighed heavily (if at all).

The present findings do not support a strict account of confirmation bias. On the contrary, the findings demonstrated that F(+I) consistently outperformed A (which itself had a higher mean viability than $A+D$ averaged across the three experiments). These findings suggest that people valued the additional information garnered by considering the tradeoff between Cells A and B. That is, they valued this particular evidential tradeoff over nonconflicting confirmation. And, they are willing to engage in a slightly more complex, integrative strategy in order to gain that additional information. The tendency to seek diagnostic information in the form of a +Itest contrast does not preclude a partial confirmation bias. Indeed, one might argue that the reason A

is weighed most heavily is that it is often associated with confirmation of a belief or hypothesis. The problem with this account is that, first, it does not explain why *D* is usually given little weight given its confirmatory status. Second, invoking the P-N distinction to account for this confirmation asymmetry is not parsimonious given that, once invoked, a P-N account does just as well on its own. Finally, in Experiments 1-3 there was no evidence of a positive correlation bias, indicating that, over each stimulus set, Cells B and C may have been viewed as no more or less confirmatory than Cells A and D. Of course, it is possible that, at other stages in the causal inference process, confirmation biases exert strong influences on judges' inferences. Those stages, however, were not the focus of this research.

Representation of P-N, I-O, and S-N as a Function of Relation Valence

Taken together, Propositions 2 and 3 state that judges tend toward +Itesting because of a primary +test bias and a secondary Stest bias. A preference for Stesting was used to explain why P-N and I-O distinctions interact. In more detailed terms, I now show that the relation between P-N and S-N is moderated by relation valence (i.e., whether the focal relation is facilitative or inhibitory) under Otest conditions only. Figure 6 shows the relations between P-N, I-O, and S-N distinctions as a function of relation valence. As can be seen in the figure, I-O is orthogonal to both P-N and S-N regardless of relation valence, and P-N and S-N are orthogonal under a facilitative set. However, under an inhibitory set, P-N and S-N correspond such that +tests are always Stests and –tests are always Ntests. The difference in the relation between P-N and S-N as a function of valence stems from the fact that the relational features of confirmation, sufficiency, and necessity change table locations by shifting cell positions within rows (i.e., Cells A and B and Cells C and D switch relational features). This has the effect that, for Itests, the relation between P-N and S-N is the same regardless of valence: +Itests = Stests and –Itests = Ntests. However, because relational features switch within row cells as a function of valence, the relation between P-N and S-N is reversed as function of valence under Otest conditions. Namely, under a facilitative set, +Otests = Ntests and –Otests = Stests, but under an inhibitory set, +Otests = Stests and –Otests = Ntests. Because P-N and S-N are orthogonal under a facilitative set but are directly related under an inhibitory set there also must be a direct relation between P-N and S-N overall across valence. In fact, when collapsed across valence, the ratio of +cell features to −cell features is 5:3 for Stesting and 3:5 for Ntesting.

Other Interpretations of Why +Itesting is so Viable

I have argued that the primary reason +Itesting is especially viable is that judges favor +testing and Stesting. The fact that judges weigh *A* most heavily and *D* least heavily indicates a primary bias toward +testing. But in order to account for the fact that judges weigh $B > C$, I posited a secondary Stest bias. Taken together, a primary +test bias and a secondary Stest bias can account for the robust SCI inequality and the inequality in test-type viabilities that follow logically (which is that the viability of +Itests > +Otests > −Otests > −Itests). At present, I know of no other account that fully and parsimoniously explains this pattern of findings. Yet I will mention two alternatives that others might view as plausible. One account is that a primary +test bias is overlaid with a secondary Itest bias. Indeed, in Proposition 4, I posited an Itest bias but noted that it must be a tertiary bias of less influence than either +test or Stest biases. The "+test and Itest bias" interpretation is lacking, however, because it does not explain why Otesting is more viable than Itesting under −test conditions. Unlike the present account, it does not explain the full pattern of results.

Another interpretation is that +Itesting outperforms +Otesting not because the former is always associated with Stesting whereas the latter may be associated with either Stesting or Ntesting, but because +Itests always provide information about the *same* criterion, which just happens to be sufficiency. This interpretation is also inadequate because, by the same argument, −Itesting should outperform −Otesting, given that the former but not the latter always provides information about necessity. −Itests, however, do not outperform −Otests. Therefore, this interpretation also fails to account for the full pattern of results.

Relations to Rule Discovery

The set of relations shown in Figure 6 represents an extension of the theoretical work by Klayman and Ha (1987) on hypothesis testing, not only because of the conceptual bridge drawn between that domain and information integration in causal inference but because of the greater generality of the present account. In their account, the relation between P-N and S-N was orthogonal because they focused primarily on the rule-discovery paradigm. The task objective there is to match a hypothesized rule to a target rule such that the former is neither overly narrow (insufficient) nor overly broad (unnecessary). Consequently, relation valence does not vary because one is always operating under a "facilitative" set (albeit there are no causal implications). Therefore, the formal representation of the rule-discovery paradigm in terms of P-N and S-N feature distinctions is a special case of the more general set of relations presented here. Only in this special case are P-N and S-N always decorrelated.

Implications for Attribution Theories

Distinguishing Variant from Invariant Cell Features

As Figure 6 illustrates, some cell features change as a function of relation valence yet others do not. Specifically, the implications of various cells in terms of the S-N distinction (encompassing the location of confirmation as well) depends on whether a focal causal relation is hypothesized or perceived to be either facilitative or inhibitory. As noted earlier, in the rule-discovery paradigm only a facilitative set is relevant, in which case the implications of cells in terms of S-N is invariant. In terms of causal judgment, however, it is incorrect to treat various cells as having invariant S-N features because either a facilitative or inhibitory set may be adopted by a judge (Kelley, 1973). Many attributional accounts (e.g., Einhorn & Hogarth, 1986; Lipe, 1991), however, have ascribed invariant S-N status to the four cells. These accounts posit that Cell B *is* the Scell, that Cell C *is* the Ncell, and that Cells A and D always define the set of confirmatory cases. On its own, this oversimplification may seem trivial (or even convenient); yet as I have shown, this treatment obscures the true nature of the relation between S-N and P-N distinctions. As new features are added (e.g., Klayman & Ha's Htest-Ttest distinction), the complexity of the relations between feature distinctions will likely increase, so it is important to establish an accurate representation of these relations. From there, special simplifying cases can always be defined.

Clarifying Zuckerman et al.'s Explanation-Inference Distinction

Zuckerman and colleagues (Zuckerman et al., 1988; Zuckerman, Eghrari, & Lambrecht, 1986) argued that attributions can represent what they term either *inferences* or *explanations*. Zuckerman et al. (1986) related these terms directly to methods for combining 2 × 2 contingency data. Accordingly, inferences involve assessing either $P(I_p|O_p)$ or $P(I_p|O_a)$ and explanations involve assessing either $P(O_p|I_p)$ or $P(O_p|I_a)$. That is, inferences correspond to Otests and explanations correspond to Itests. They suggest further that the +Otests and +Itests are the "core elements in the inference and explanation sets, respectively" (p. 1144); however, they do not state why this would be so. A problem with their account is that they equate +Otesting with Ntesting and +Itesting with Stesting, concluding that "it seems likely that the concepts of necessity and sufficiency are related to those of inference and explanation, respectively" (p. 1151). However, as Figure 6 shows, +Otesting, like –Otesting, can be associated with either Ntesting or Stesting. Moreover, –Itesting (related to Zuckerman et al.'s explanation set) is consistently an Ntest, not an Stest as the last quote suggests.

Forward versus Backward Causal Inference

The present findings support the notion that judges focus more on Stesting than Ntesting. This account, however, is in apparent contradiction with many recent accounts of causal judgment (e.g., Einhorn & Hogarth, 1986; Hilton, 1990; Kahneman & Miller, 1986; Lipe, 1991; McGill & Klein, 1993) that posit an important, if not dominant, role of Ntesting in arriving at causal inferences. These accounts owe much to Mackie's (1974; see also Hart & Honoré, 1959) philosophical account of causation. Mackie noted that in judging whether Antecedent X (facilitatively) caused Outcome Y, one asks the counterfactual question, "Would Y have happened if X had not?" Using the present terminology, this question translates into "$P(O_p|I_a)$?" This suggests that judges might contrast Cells C and D as in –Itesting, which always represents an Ntest (see Figure 6).

Einhorn and Hogarth (1986) termed this type of counterfactual reasoning *backward causal inference* (BCI) because the judge already knows *de facto* that the outcome has occurred and is presumably trying to reason back from this outcome to a possible cause. In contrast, in *forward causal inference* (FCI) the outcome is unknown and the question of interest is "Will X lead to Y?" Einhorn and Hogarth suggested that BCIs focus on necessity (or what they term the multiplicity of causes), whereas FCIs focus on sufficiency (or what they term the conditionality of causes). However, there are problems with linking the direction of causal inference to the S-N distinction in this manner. And, it is unlikely that everyday counterfactual conditionals represent tests of causal necessity. Next, I discuss each of these issues in turn.

Causal Inference Direction: Isomorph of S-N or I-O?

If BCI entails reasoning from known outcomes to hypothesized causes, then one might expect it to have a closer association with the I-O distinction than with the S-N distinction. For example, a BCI might take the form of the question, "Given that *O* occurred, how likely is it that I_p was the cause?" In contrast to the counterfactual question, which takes the form of a –Itest, this BCI question takes the form of a +Otest. And, whereas the counterfactual question always provides information on necessity, the BCI question provides information on necessity only under a facilitative set. Alternatively, an FCI question might take the form, "Given that *I* occurred, how likely is it that O_p will follow?" This question corresponds to a +Itest, which always provides information on sufficiency. Thus, FCI might correspond with Itesting (especially +Itesting) and BCI might correspond with Otesting (especially +Otesting; see Zuckerman et al., 1986, 1988, for a similar account). This interpretation coheres with the present account because both FCI and BCI would favor +testing.

Everyday Counterfactual Conditionals

Mandel and Lehman (in press) have recently challenged the notion that everyday counterfactual conditionals represent tests of hypothesized causal relations. They noted that such counterfactuals are most common following affectively negative outcomes (see also Taylor, 1991) and often take the form, "If only X hadn't happened, then Y wouldn't have happened either." These statements often have a confirmatory rather than test-like quality. If counterfactual conditionals represented causal questions of a –Itest form in which the data were mentally simulated (Kahneman & Tversky, 1982), then one might expect that, on average across individuals and situations, a sizable proportion of the "answers" would represent C-cell instances. Although no study has examined the proportion of spontaneous counterfactual conditionals that correspond to either C-cell or D-cell instances, intuition suggests that the vast majority represent D-cell instances. In contrast to the "Ntest interpretation," Mandel and Lehman (in press) argued that these counterfactual D-cell instances represent examples that come to mind of how an affectively negative outcome could have been prevented. Consider, for instance, the counterfactual statement, "If only I hadn't bothered to check my mail I wouldn't have missed my bus." The standard interpretation is that the statement represents a confirmation of the hypothesis that, on that occasion, "checking my mail was a necessary cause of missing my bus." Mandel and Lehman argued instead that the statement represents a confirmation of the hypothesis that, on that occasion, "not checking my mail would have been sufficient to prevent missing the bus."

The implication of the Mandel and Lehman interpretation is that even in situations in which people generate everyday counterfactual conditionals of the "if only . . ." form, their primary concern is not with necessity but with sufficiency. In logical terms, counterfactual conditionals have implications for both necessary facilitative causes and sufficient preventors. In psychological terms, however, there appears to be a closer connection between counterfactuals and thoughts about sufficient preventors. As noted earlier, counterfactual conditionals are especially common following affectively negative outcomes. In these situations, people often are concerned with identifying controllable actions that they (or some other focal actor) could have taken that would have been sufficient to prevent the outcome from happening. Supporting this interpretation, Mandel and Lehman found that when subjects were given vignettes of situations in which affectively negative outcomes (e.g., a plane crash resulting in the death of a focal actor) occurred, subjects' judgments of the cause of the outcome focused on events that general world knowledge indicates would covary with the target outcome (e.g., engine malfunctions). However, both subjects' judgments of how the outcome could have been prevented and their judgments of what a focal actor would have thought counterfactually tended to focus on controllable actions that the focal actor could have taken (e.g., choosing to take the train rather than to fly). These findings have interesting implications for the present account because they suggest that counterfactual conditional reasoning may still represent a focus on sufficiency, albeit one that concerns perceived controllable preventors that might (or could) have been enacted prior to the cause of an outcome. A further implication is that counterfactual reasoning may represent one domain in which negational thinking dominates over affirmational thinking resulting in a focus on Cell D. This makes sense if the counterfactual case is sufficient to cognitively undo the actual case which, in the first place, was (and remains) likely to be viewed in affirmational terms.

So What About Ntesting?

Although I have argued that causal inference is biased toward Stesting and have refuted some of the main arguments in favor of necessity theories of causation, I do not deny the important role that Ntesting can play in certain types of causal thinking. Clearly, the *cause in fact* or *but for cause* that figures prominently in legal reasoning is an example of a situation in which Ntesting is an important focus of causal thought (e.g., see Hart & Honoré, 1959). For example, one might state counterfactually that had Jeffery Dahlmer's parents never met, Dahlmer's victims would still be alive today. Although most people would not view the meeting of Dahlmer's parents as a cause of the victims' deaths, that factor nevertheless constitutes a *but for cause* because the negation of that factor would have undone the target event. In Cheng and Novick's (1991, 1992) terminology, the meeting of Dahlmer's parents would represent an enabling condition rather than a cause, much like oxygen is an enabling condition (i.e., it is necessary) for fire to occur but would not be viewed as the cause of fire under normal circumstances (viz., within the focal set of cases in which oxygen is always present). As already noted, in psychological terms, much of people's everyday counterfactual thinking might not even be concerned with the isolation of such enabling conditions.

Why might necessity be associated with mere enabling conditions whereas sufficiency appears to be associated with causation? One possibility is that the number of necessary conditions for an outcome is overwhelmingly large: For just about any event, we could (counterfactually) trace back through time subtle changes that might have disabled the event. However, most of these enabling conditions are psychologically meaningless. The overabundance of necessary constraints relegates them to the status of mere enabling conditions. In contrast, the set of conditions that are sufficient to produce an outcome is typically relatively small. And, whereas one can mentally simulate moving back a step in time to conjure up additional enabling (or disabling) conditions, the same cannot be said for sufficient conditions, which generally appear to be in relatively close temporal proximity to target outcomes. The scarcity of sufficient conditions relative to necessary ones, therefore, may contribute to the former generally having *causal* status whereas the latter generally has *condition* status.

Limitations and Future Directions for Research

Experimental Characteristics

The nature of the judgment task in the present three experiments is representative in many ways of the type of research conducted in this area. These similarities were intentional, given the fact that I have proposed a new framework for this type of research. Nevertheless, there are several aspects of the research that do deserve comment. One issue concerns the manner in which subjects receive information in the present research. Clearly, outside of the lab judges do not receive contingency data in a neatly summarized manner in which exact numeric frequencies are provided. This procedure was adopted purposely in order to minimize intrusions of other cognitive demands. And, steps were taken to compare the results of the present experiments to independent studies that have used modified procedures such as discrete-trial data presentations.

Still it is worth considering whether or not the data-processing context that exists in the lab is categorically different from that which exists outside the lab. For instance, the mere exposure to *numeric* data may prompt subjects to arrive at causal inferences in ways that would not otherwise be likely. If such were the case, not only the present account but much of research on covariational assessment and causal inference would be called into question. Perhaps, then, this type of research has more to do with what Busemeyer (1991) called intuitive statistical estimation than causal inference per se. My view is that this research intersects both of these descriptions. One possible direction for future research is to examine judges' processing of data that is precise but non-numeric. For example, I know of no covariation study that has represented the frequencies of event conjunctions as areas or lengths rather than as numbers. Doing so would allow for precise rule predictions (because areas can be translated into numbers) without forcing numbers on subjects. A group of subjects might also estimate the percentage of total area occupied by each conjunction (i.e., cell) to get an estimate of their perception of non-numeric magnitude.

In Experiment 3, I ruled out an information-primacy account of the obtained SCI inequality and I examined the effect of event type (abstract vs. content-based) on SCI. Despite these attempts to clarify the meaning of the present findings, in hindsight, other precautions could have been taken and might be usefully incorporated in future research. For example, the instructions in the present experiments focus on characterizing the effect of Event A on Event B. These instructions focus explicitly on the positive state of event presence and may have contributed to the strength of the observed +test bias. Alternatively, one might ask subjects to characterize the effect of the presence or absence of Event A on the presence or absence of Event B. Although these instructions are unbiased in terms of the P-N distinction, they sound odd and cumbersome because statements *are* typically phrased in the affirmative (Clark & Clark, 1977). Moreover, Crocker (1982) found that when this type of unbiased question wording was used, 100% of subjects indicated that A-cell data were necessary for assessing covariation, whereas only 35% of subjects indicated that D-cell data were necessary. This finding indicates that judges weigh positive information more heavily than negative information even when P-N-unbiased instructions are used.

A related concern is that the present instructions focused on the effect of Event A on Event B. This may have prompted subjects to focus on FCI, which may have contributed to the observed Itest bias. Thus, in future research the emphasis on either FCI or BCI may be directly manipulated through question wording. The ordering of cell frequencies also may affect preferences for Itesting or Otesting. That is, in the present experiments Cells A and B and Cells C and D always were adjacent. This ordering is consistent with +Itests and –Itests, respectively. Presenting data in the A-C-B-D order might have increased the viability of Otesting. In a related vein, the manner in which event conjunctions are described could be manipulated in future research. For example, Cell A was described as, "Event A occurred and then Event B occurred." It may alternatively be described in the present research as, "Event B occurred and was preceded by Event A." This manipulation may induce different temporal perspectives (i.e., FCI vs. BCI) and may address possible event-order effects (e.g., Kahneman & Tversky's, 1982, focus rule).

The Present Account

The present account of information integration in causal inference was supported by the findings of this research. Nevertheless, the account may be extended in various ways. For one thing, the present research focused on the simplest case in which there was only one antecedent and one consequent binary variable. As the previous suggestions for future research make clear, even within the context of this simple case much research is needed. At the same time, extensions of the account to include the simultaneous processing of multiple inputs and outcomes will eventually be required. Examining the multiplicity of inputs, for example, is not just important because it more closely reflects conditions in the real world, it is also important for theoretical reasons that relate directly to propositions of the present account. For example, Einhorn and Hogarth (1986) suggested that a focus on necessity concerns the issue of causal multiplicity (i.e., can other causes yield the effect?), whereas a focus on sufficiency concerns the issue of causal conditionality (i.e., are there conditions that limit the ability of a cause to yield the effect? e.g., the absence of an enabling condition or the presence of a disabling condition). Therefore, it is important to examine the effects of adding not just more inputs to a causal scenario but different types of inputs. If Einhorn and Hogarth are correct, one might find that adding additional potential causes increases a judge's focus on Ntesting. Alternatively, manipulating the presence or absence of various enabling or disabling conditions may strengthen a judge's focus on Stesting.

Another important direction for future research will be to examine how judgments of relation strength translate into different types of causal judgments. Presently, we know little about the types of situations in which causal relations are defined either dichotomously (i.e., causal vs. noncausal) or along a continuum (i.e., noncausal; weak, moderate, or strong causal relation). In almost all covariational research (including the present experiments), subjects are forced to respond in terms of a causal/covariational continuum. This imposition is convenient given that the rules examined in such research yield values along a continuum, yet the correspondence of such judgments to natural categories of causal judgment needs to be explored. Again, the implications of this issue extend beyond one of ecological validity to theoretical concerns. For instance, if causal relations are viewed along a strength continuum, then potential causes might interact in a compensatory nature (Kelley, 1971). Either of two strong causes may be sufficient to yield an effect, but one moderate cause might require an additional moderate cause in order to be sufficient to yield the effect.

Proposition 3 of the present account was articulated in terms of P-N and I-O feature distinctions. However, the interpretation of the proposed P-N × I-O interaction was explained in terms of the S-N distinction. Because the representation of S-N changes as a function of relation valence under Otest conditions, future research might test the predictions of the present account by examining judgments of facilitative and inhibitory causation separately. If the present account is correct, ratings of facilitative causal strength should replicate the typical SCI inequality. However, ratings of inhibitory causal strength should result in judges weighing $A > B \cong C > D$ because Cells B and C represent confirmatory cases. That is, Cell B is no longer uniquely associated with Stesting under an inhibitory set. In effect, the $B > C$ weighting inequality would be transferred to the $A > D$ weighting inequality in a manner that amplifies the latter inequality. An additional predicted effect, then, is that the range of weight values is more compressive under facilitative sets than under inhibitory sets.

CONCLUSION

The logic and methodology of the present feature-analytic framework for research on information integration in causal inference represents an advancement that builds on the extant tradition of rule-analytic research. The descriptive account developed within this new framework posits that judges are (a) concerned with evidential tradeoffs, leading to a reliance on contrastive strategies (Proposition 1), (b) primarily biased toward inferential strategies that focus on positive information (Proposition 2), (c) secondarily biased toward assessing sufficiency rather than necessity, as indicated by the predicted P-N × I-O interaction (Proposition 3), and (d) somewhat biased toward strategies that cohere with our everyday sense of time flowing from causes to effects (viz., Itesting; Proposition 4). The account also posits that judges tend toward probability strategies that allow for comparability across data contexts rather than frequency strategies that do not (Proposition 5). Empirical tests of the account strongly supported Propositions 1-4 and tentatively supported Proposition 5. More generally, the findings supported the view that invariant properties of information integration processes may best be captured at the level of features rather than at the level of rules.

REFERENCES

Abelson, R. P., & Levi, A. (1985). Decision making and decision theory. In G. Lindzey & E. Aronson, (Eds.), *Handbook of experimental social psychology* (Vol. 2, pp. 231-310). New York: Random House.

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.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120,* 3-19.

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.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

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

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

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

Bassok, M., & Trope, Y. (1984). People's strategies for testing hypotheses about another's personality: Confirmatory or diagnostic? *Social Cognition, 2,* 199-216.

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

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition, 14,* 237-273.

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.

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

Cheng, P. W. (1993). Separating causal laws from causal facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 30, pp. 215-264). New York: Academic Press.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293-328.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40,* 83-120.

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 Inc.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), The adapted mind: Evolutionary psychology and the generation of culture (pp. 117-182). New York: Oxford University Press.

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.

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

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

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.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239-260.

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.

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

Fratianne, A., & Cheng, P. W. (1995). *Assessing causal relations by dynamic hypothesis testing.* Manuscript submitted for publication.

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

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry, 5,* 459-464.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3: Speech acts* (pp. 41-58). New York: Academic Press.

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

Hart, H. L., & Honoré, A. M. (1959). *Causation in the law.* Oxford, England: Clarendon Press.

Hearst, E. (1978). Stimulus relationships and feature selection in learning and behavior. In S. Hulse, H. Fowler, & W. K. Honig (Eds.), *Cognitive processes in animal behavior.* Hillsdale, NJ: Erlbaum.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107,* 65-81.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review, 93,* 75-88.

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.

Hume, D. (1938). *An abstract of a treatise of human nature.* London: Cambridge University Press. (Original work published 1740)

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* London: Routledge and Kegan Paul.

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

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., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136-153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201-208). New York: Cambridge University Press.

Kanazawa, S. (1992). Outcome or expectancy? Antecedent of spontaneous causal attribution. *Personality and Social Psychology Bulletin, 18,* 659-668.

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

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.

Kasof, J., & Lee, J. Y. (1993). Implicit causality as implicit salience. *Journal of Personality and Social Psychology, 65,* 877-891.

Kelley, H. H. (1971). Causal schemata and the attribution process. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 151-174). Morristown, NJ: General Learning Press.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28,* 103-128.

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.

Krosnick, J. A., Li, F., & Lehman, D. R. (1990). Conversational conventions, order of information acquisition,, and the effect of base rates and individuating information on social judgments. *Journal of Personality and Social Psychology, 59,* 1140-1152.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480-498.

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

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

Lipe, M. G. (1991). Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin, 109,* 456-471.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47,* 1231-1243.

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

Mandel, D. R., & Lehman, D. R. (in press). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology.*

McGill, A. L. (1989). Context effects in judgments of causation. *Journal of Personality and Social Psychology, 57,* 189-200.

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

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.

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 on mental processes. *Psychological Review, 84,* 231-259.

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

Pyszczynski, T. A., & Greenberg, J. (1981). Role of disconfirmed expectancies in the instigation of attributional processing. *Journal of Personality and Social Psychology, 40,* 31-38.

Read, S. J. (1987). Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology, 52,* 288-302.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General, 122,* 92-114.

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

Sanbonmatsu, D. M., Akimoto, S. A., & Biggs, E. (1993). Overestimating causality: Attributional effects of confirmatory processes. *Journal of Personality and Social Psychology, 65,* 892-903.

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

Shaklee, H., & Mims, M. (1982). Sources of error in judging event 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. (1985). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition, 13,* 158-167.

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

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

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

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22,* 93-121.

Slovic, P., Lichtenstein, S., & Fischhoff, B. (1985). Decision making. In R. C. Atkinson, R. J. Herrnstein, G. D. Lindzey, & R. D. Luce (Eds.), *Handbook of experimental psychology* (Vol. 2, pp. 673-738). New York: Wiley.

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.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin, 110,* 67-85.

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

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology, 43,* 22-34.

Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology, 19,* 560-576.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

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

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.

Wason, P. C. (1968). On the failure to eliminate hypotheses—A second look. In P. C. Wason & P. N. Johnson-Laird (Eds.), *Thinking and reasoning* (pp. 165-174). Harmondsworth, Middlesex, England: Penguin.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* London: Batsford.

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

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

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

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.

Weiner, B. (1985). "Spontaneous" causal thinking. *Psychological Bulletin, 97,* 74-84.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56,* 161-169.

Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science, 5,* 249-254.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116,* 117-142.

Wong, P. T. P., & Weiner, B. (1981). When people ask "why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology, 40,* 650-663.

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

Zuckerman, M., Colwell, E. L., Darche, P. R., Fischer, S. A., Osmun, R. W., Spring, D. D., Winkler, B. A., & Wolfson, L. R. (1988). Attributions as inferences and explanations: Effects on discounting. *Journal of Personality and Social Psychology, 54,* 748-757.

Zuckerman, M., Eghrari, H., & Lambrecht, M. R. (1986). Attributions as inferences and explanations: Conjunction effects. *Journal of Personality and Social Psychology, 51,* 1144-1153.

FOOTNOTES

[1] Some literature (e.g., Allen & Brooks, 1991; Regehr & Brooks, 1993) on perceptual integration uses the term *feature* to refer to attributes of sets of stimuli that subjects perceive rather than attributes of sets of rules that may be used to model subjects' responses.

[2] It is difficult to determine, however, whether changes in strategy reflect changes in the perceptual salience of different variables, the complexity of implementing a particular strategy given the type of data presented, or both factors.

[3] Note that an extension of the confirming-cases strategy to joint probabilities is contrastive rather than cumulative. That is, computing $P(A \cup D)$, namely $(A + D)/N$, provides information about the tradeoff between CFAC cases and CINH cases. This is because $P(A \cup D) + P(B \cup C) = 1$ (a constant). In contrast, computing $f(A \cup D)$ does not automatically inform a judge of $f(B \cup C)$ because these two frequencies sum to the variable N. (The importance of calculating joint probabilities rather than frequencies is discussed further in a later section.)

[4] Consistent with the views of other researchers (e.g., Einhorn & Hogarth, 1986; Fratianne & Cheng, 1996; Jenkins & Ward, 1965; Schustack & Sternberg, 1981; Suppes, 1970), it is assumed that necessity and sufficiency criteria are generally evaluated in probabilistic rather than deterministic terms. In facilitative terms, necessity is undermined as C increases relative to A and/or D, and sufficiency is undermined as B increases relative to A and/or D. In inhibitory terms, necessity is undermined as D increases relative to B and/or C, and sufficiency is undermined as A increases relative to B and/or D.

[5] There are sixteen combinatorial patterns of high and low values across the four conditional probabilities (i.e., 4C0H = 1, 4C1H = 4, 4C2H = 6, 4C3H = 4, and 4C4H = 1). However, due to the fact that information in the numerators of Pc(+I) and Pc(–O) and Pc(–I) and Pc(+O) is redundant, two patterns (viz., LLHH and HHLL) are impossible to construct.

[6] Experiments 2 and 3 by Wasserman et al. (1990) used the same stimuli and method for presenting data. Given this fact, and the fact that the mean causal ratings of subjects in these experiments correlated .998, the ratings for the two experiments were averaged.

Table 1

Notation and Computations for Rules Modeling P-N, I-O, and F-P Feature Conjunctions

	Type of Information that is Integrated			
	+test		−test	
How?[a]	Itest	Otest	Itest	Otest
Ftest	F(+I) = $A - B$	F(+O) = $A - C$	F(−I) = $D - C$	F(−O) = $D - B$
Pjtest	Pj(+I) = $(A - B)/N$	Pj(+O) = $(A - C)/N$	Pj(−I) = $(D - C)/N$	Pj(−O) = $(D - B)/N$
Pctest	Pc(+I) = $A/(A + B)$	Pc(+O) = $A/(A + C)$	Pc(−I) = $D/(D + C)$	Pc(−O) = $D/(D + B)$

[a]How is the information integrated?

Table 2

Stimuli and Mean Causal Ratings in Experiment 1

		Cell Frequency				Marg. *P*		Ratings	
Stim.	Pattern	*A*	*B*	*C*	*D*	I_p	O_p	*M*	*SD*
1	LLLL	47,500	2,500	2,500	47,500	.50	.50	.56	.42
2	LLLH	94,715	4,985	15	285	1	.95	.76	.38
3	LLHL	285	15	4,985	94,715	.00	.05	–.03	.55
4	LHLL	94,715	15	4,985	285	.95	1	.72	.35
5	LHLH	90,250	4,750	4,750	250	.95	.95	.57	.36
6	LHHL	4,750	250	90,250	4,750	.05	.95	–.20	.49
7	LHHH	4,985	285	94,715	15	.05	1	–.17	.52
8	HLLL	285	4,985	15	94,715	.05	.00	–.08	.59
9	HLLH	4,750	90,250	250	4,750	.95	.05	–.62	.45
10	HLHL	250	4,750	4,750	90,250	.05	.05	–.04	.46
11	HLHH	15	94,715	285	4,985	.95	.00	–.69	.45
12	HHLH	4,985	94,715	285	15	1	.05	–.54	.46
13	HHHL	15	285	94,715	4,985	.00	.95	–.24	.56
14	HHHH	2,500	47,500	47,500	2,500	.50	.50	–.21	.39

Note. Stim. = stimulus number. Pattern refers to high (H) or low (L) values for Pc(+I), Pc(–I), Pc(+O), and Pc(–O), respectively. Marg. *P* = marginal probability.

Table 3

Stimuli and Mean Causal Ratings in Experiment 2

	Cell Frequency				Marg. *P*		Pctest				Ratings	
Stim.	*A*	*B*	*C*	*D*	I_p	O_p	+I	–I	+O	–O	*M*	*SD*
1	4	0	0	0	1	1	0	●	0	●	.90	.28
2	0	4	0	0	1	0	1	●	●	1	-.24	.18
3	0	0	4	0	0	1	●	1	1	●	-.05	.22
4	0	0	0	4	0	0	●	0	●	0	.08	.18
5	3	1	0	0	1	.75	.25	●	0	1	.20	.12
6	3	0	1	0	.75	1	0	1	.25	●	.20	.10
7	3	0	0	1	.75	.75	0	0	0	0	.29	.08
8	1	3	0	0	1	.25	.75	●	0	1	-.14	.16
9	0	3	1	0	.75	.25	1	1	1	1	-.21	.15
10	0	3	0	1	.75	0	1	0	●	.75	-.15	.15
11	1	0	3	0	.25	1	0	1	.75	●	-.03	.15
12	0	1	3	0	.25	.75	1	1	1	1	-.15	.17
13	0	0	3	1	0	.75	●	.75	1	0	-.08	.14
14	1	0	0	3	.25	.25	0	0	0	0	.17	.17
15	0	1	0	3	.25	0	1	0	●	.25	.01	.14
16	0	0	1	3	0	.25	●	.25	1	0	.03	.13
17	2	1	1	0	.75	.75	.33	1	.33	1	.09	.11
18	2	1	0	1	.75	.50	.33	0	0	.50	.17	.12
19	2	0	1	1	.50	.75	0	.50	.33	0	.13	.14
20	1	2	1	0	.75	.50	.67	1	.50	1	-.10	.14
21	1	2	0	1	.75	.25	.67	0	0	.67	-.05	.13
22	0	2	1	1	.50	.25	1	.50	1	.67	-.14	.16

(table continues)

23	1	1	2	0	.50	.75	.50	1	.67	1	-.03	.12
24	1	0	2	1	.25	.75	0	.67	.67	0	-.01	.12
25	0	1	2	1	.25	.50	1	.67	1	.50	-.09	.15
26	1	1	0	2	.50	.25	.50	0	0	.33	.10	.13
27	1	0	1	2	.25	.50	0	.33	.50	0	.10	.13
28	0	1	1	2	.25	.25	1	.33	1	.33	-.01	.14
29	2	2	0	0	1	.50	.50	●	0	1	.06	.11
30	2	0	2	0	.50	1	0	1	.50	●	.02	.09
31	2	0	0	2	.50	.50	0	0	0	0	.25	.12
32	0	2	2	0	.50	.50	1	1	1	1	-.20	.17
33	0	2	0	2	.50	0	1	0	●	.50	-.03	.12
34	0	0	2	2	0	.50	●	.50	1	0	-.02	.10
35	1	1	1	1	.50	.50	.50	.50	.50	.50	.01	.06

Note. Stim. = stimulus number. Marg. *P* = marginal probability. ● = undefined value (viz. 0/0).

Table 4

Stimuli and Mean Causal Ratings in Experiment 3

	Cell Frequency				Marg. *P*		Pctest				Pjtest				Ratings	
Dist.	*A*	*B*	*C*	*D*	I_p	O_p	+I	–I	+O	–O	+I	–I	+O	–O	*M*	*SD*
N = 10																
1	**5**	2	2	1	.7	.7	.29	.67	.29	.67	.5	.2	.2	.1	.48	.33
2	**5**	2	1	2	.7	.6	.29	.33	.17	.50	.5	.2	.1	.2	.52	.33
3	**5**	1	2	2	.6	.7	.17	.50	.29	.33	.5	.1	.2	.2	.58	.26
4	1	**5**	2	2	.6	.3	.83	.50	.67	.71	.1	.5	.2	.2	–.44	.38
5	2	**5**	2	1	.7	.4	.71	.67	.50	.83	.2	.5	.2	.1	–.44	.36
6	2	**5**	1	2	.7	.3	.71	.33	.33	.71	.2	.5	.1	.2	–.31	.33
7	1	2	**5**	2	.3	.6	.67	.71	.83	.50	.1	.2	.5	.2	–.31	.40
8	2	1	**5**	2	.3	.7	.33	.71	.71	.33	.2	.1	5	.2	–.11	.34
9	2	2	**5**	1	.4	.7	.50	.83	.71	.67	.2	.2	.5	.1	–.19	.40
10	2	2	1	**5**	.4	.3	.50	.17	.33	.29	.2	.2	.1	.5	.12	.42
11	2	1	2	**5**	.3	.4	.33	.29	.50	.17	.2	.1	.2	.5	.08	.41
12	1	2	2	**5**	.3	.3	.67	.29	.67	.29	.1	.2	.2	.5	–.04	.34
N = 20																
1	**10**	4	4	2	.7	.7	.29	.67	.29	.67	.5	.2	.2	.1	.51	.28
2	**10**	4	2	4	.7	.6	.29	.33	.17	.50	.5	.2	.1	.2	.62	.31
3	**10**	2	4	4	.6	.7	.17	.50	.29	.33	.5	.1	.2	.2	.62	.26
4	2	**10**	4	4	.6	.3	.83	.50	.67	.71	.1	.5	.2	.2	–.62	.31
5	4	**10**	4	2	.7	.4	.71	.67	.50	.83	.2	.5	.2	.1	–.47	.40
6	4	**10**	2	4	.7	.3	.71	.33	.33	.71	.2	.5	.1	.2	–.38	.40

(table continues)

7	2	4	**10**	4	.3	.6	.67	.71	.83	.50	.1	.2	.5	2	−.35	.46
8	4	2	**10**	4	.3	.7	.33	.71	.71	.33	.2	.1	5	.2	−.15	.43
9	4	4	**10**	2	.4	.7	.50	.83	.71	.67	.2	.2	.5	.1	−.09	.50
10	4	4	2	**10**	.4	.3	.50	.17	.33	.29	.2	.2	.1	.5	.12	.44
11	4	2	4	**10**	.3	.4	.33	.29	.50	.17	.2	.1	.2	.5	.18	.44
12	2	4	4	**10**	.3	.3	.67	.29	.67	.29	.1	.2	.2	.5	−.06	.41
								N = 40								
1	**20**	8	8	4	.7	.7	.29	.67	.29	.67	.5	.2	.2	.1	.59	.38
2	**20**	8	4	8	.7	.6	.29	.33	.17	.50	.5	.2	.1	.2	.73	.30
3	**20**	4	8	8	.6	.7	.17	.50	.29	.33	.5	.1	.2	.2	.74	.29
4	4	**20**	8	8	.6	.3	.83	.50	.67	.71	.1	.5	.2	.2	−.68	.39
5	8	**20**	8	4	.7	.4	.71	.67	.50	.83	.2	.5	.2	.1	−.66	.36
6	8	**20**	4	8	.7	.3	.71	.33	.33	.71	.2	.5	.1	.2	−.50	.41
7	4	8	**20**	8	.3	.6	.67	.71	.83	.50	.1	.2	.5	2	−.34	.29
8	8	4	**20**	8	.3	.7	.33	.71	.71	.33	.2	.1	5	.2	−.14	.50
9	8	8	**20**	4	.4	.7	.50	.83	.71	.67	.2	.2	.5	.1	−.23	.50
10	8	8	4	**20**	.4	.3	.50	.17	.33	.29	.2	.2	.1	.5	.13	.50
11	8	4	8	**20**	.3	.4	.33	.29	.50	.17	.2	.1	.2	.5	.19	.53
12	4	8	8	**20**	.3	.3	.67	.29	.67	.29	.1	.2	.2	.5	−.09	.42

Note. Dist. = frequency distribution. Marg. *P* = marginal probability. Bolded values denote stimulus subsets within each set size whose largest frequency is either *A*, *B*, *C*, or *D*.

Table 5

Mean Viabilities (r') of Cell Frequencies in Experiments 1-3

	Experiment			
Frequency	1 (*n* = 52)	2 (*n* = 51)	3 (*n* = 120)	Weighted *M*
A	.82 (.32)	.70 (.23)	.50 (.15)	.62 (.21)
B	−.61 (.31)	−.56 (.26)	−.41 (.17)	−.49 (.22)
C	−.17 (.33)	−.30 (.24)	−.09 (.24)	−.16 (.26)
D	.04 (.30)	.19 (.19)	.04 (.21)	.07 (.23)

Note. Values in parentheses are standard deviations.

Table 6

Percentage of Cell-Frequency Viabilities as a Function of Cell-Frequency Rank

	Cell-Frequency Viability			
Rank	*A*	*B*	*C*	*D*
	Experiment 1			
1	***67***	35	11	4
2	21	***44***	17	23
3	8	11	***37***	35
4	4	10	35	***38***
	Experiment 2			
1	***65***	27	10	0
2	29	***53***	22	4
3	6	14	***43***	37
4	0	6	25	***59***
	Experiment 3			
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% due to ties that increase the proportion of higher ranks. Bolded values identify the most frequent ranking for each cell-frequency viability. Italicized values identify the cell-frequency viability that is most frequent for a given rank.

Table 7

Mean Viabilities (r') of Ftests as a Function of P-N and I-O Distinctions in Experiments 1-3

	P-N		
I-O	+test	−test	$M_{\text{I-O}}$
Experiment 1			
Itest	.97 (.42)	.13 (.33)	.55 (.38)
Otest	.57 (.31)	.37 (.30)	.47 (.31)
$M_{\text{P-N}}$	.77 (.37)	.25 (.32)	.51 (.35)
Experiment 2			
Itest	.85 (.34)	.30 (.24)	.58 (.29)
Otest	.61 (.24)	.45 (.23)	.53 (.24)
$M_{\text{P-N}}$	.73 (.29)	.38 (.24)	.56 (.27)
Experiment 3			
Itest	.81 (.23)	.11 (.32)	.46 (.28)
Otest	.47 (.26)	.36 (.24)	.42 (.25)
$M_{\text{P-N}}$	.64 (.25)	.24 (.28)	.44 (.27)
Weighted *M* for Experiments 1-3			
Itest	.86 (.30)	.16 (.30)	.51 (.30)
Otest	.53 (.27)	.38 (.25)	.46 (.26)
$M_{\text{P-N}}$	.70 (.29)	.27 (.28)	.49 (.29)

Note. Values in parentheses are standard deviations.

Table 8

Viabilities as a Function of P-N and I-O Distinctions in Research Using Discrete-trial Methods for Presenting Data

	P-N		
I-O	+test	−test	$M_{\text{I-O}}$
	Anderson and Sheu (1995, Exp. 1)		
	$N_{\text{stimuli}} = 80$, $N_{\text{subject}} = 40$		
Itest	1.29	.34	.82
Otest	.87	.62	.75
$M_{\text{P-N}}$	1.08	.48	.79
	Kao and Wasserman (1993, Exp. 2)		
	$N_{\text{stimuli}} = 21$, $N_{\text{subject}} = 99$		
Itest	.94	−.23	.36
Otest	.40	.08	.24
$M_{\text{P-N}}$	.67	−.07	.30

Note. The data from Kao and Wasserman is based on only the subjects in the discrete-trial condition. Viabilities are based on the average of Ftest, Pctest, and Pjtest measures (i.e., they are collapsed across the F-P distinction).

Table 9

Viabilities as a Function of the F-P Distinction in Six Experiments

			F-P			
Experiment	N_{stim}	N_{subj}	F	Pc	Pj	*M*
Mandel (1996, Exp. 3) Across Event Type[a]	36	120	.44	.46	.49	.46
			(.20)	(.21)	(.22)	(.21)
Abstract	36	40	.46	.45	.50	.47
			(.24)	(.25)	(.28)	(.26)
Content-Natural	36	40	.48	.52	.55	.52
			(.14)	(.17)	(.16)	(.16)
Content-Social	36	40	.37	.39	.43	.40
			(.17)	(.19)	(.20)	(.19)
Anderson & Sheu (1995, Exp. 1)	80	40	.78	.87	.82	.82
Kao & Wasserman (1993, Exp. 2, trial)	21	99	.36	.23	.31	.30
Kao & Wasserman (1993, Exp. 2, summary)	21	101	.35	.25	.31	.31
Levin et al. (1993, Exp. 1)	31	80	.80	.85	.84	.83
Wasserman et al. (1990, Exp. 2 and 3)	25	240	1.30	1.41	1.30	1.34

Note. N_{stim} = number of stimuli, N_{subj} = number of subjects.

[a]The values presented are mean viabilities and standard deviations are presented in parentheses.

	O_p	O_a	
I_p	**Cell A** $A = f(I_p \cap O_p)$	**Cell B** $B = f(I_p \cap O_a)$	$A + B = f(I_p)$
I_a	**Cell C** $C = f(I_a \cap O_p)$	**Cell D** $D = f(I_a \cap O_a)$	$C + D = f(I_a)$
	$A + C = f(O_p)$	$B + D = f(O_a)$	$A + B + C + D = N$

Figure 1. The standard 2 × 2 contingency table wherein a causal input (*I*) is either present (I_p) or absent (I_a) and a causal output (*O*) is either present (O_p) or absent (O_a). The letters *A* to *D* correspond to particular cell frequencies.

Summary #1:

Observation	**Number of instances out of 100,000**
Event A OCCURRED and then Event B OCCURRED	47,500
Event A OCCURRED and then Event B DID NOT OCCUR	2,500
Event A DID NOT OCCUR and then Event B OCCURRED	2,500
Event A DID NOT OCCUR and then Event B DID NOT OCCUR	47,500

Figure 2. An example taken from Experiment 1 of the format used in Experiments 1-3 for presenting contingency information to subjects.

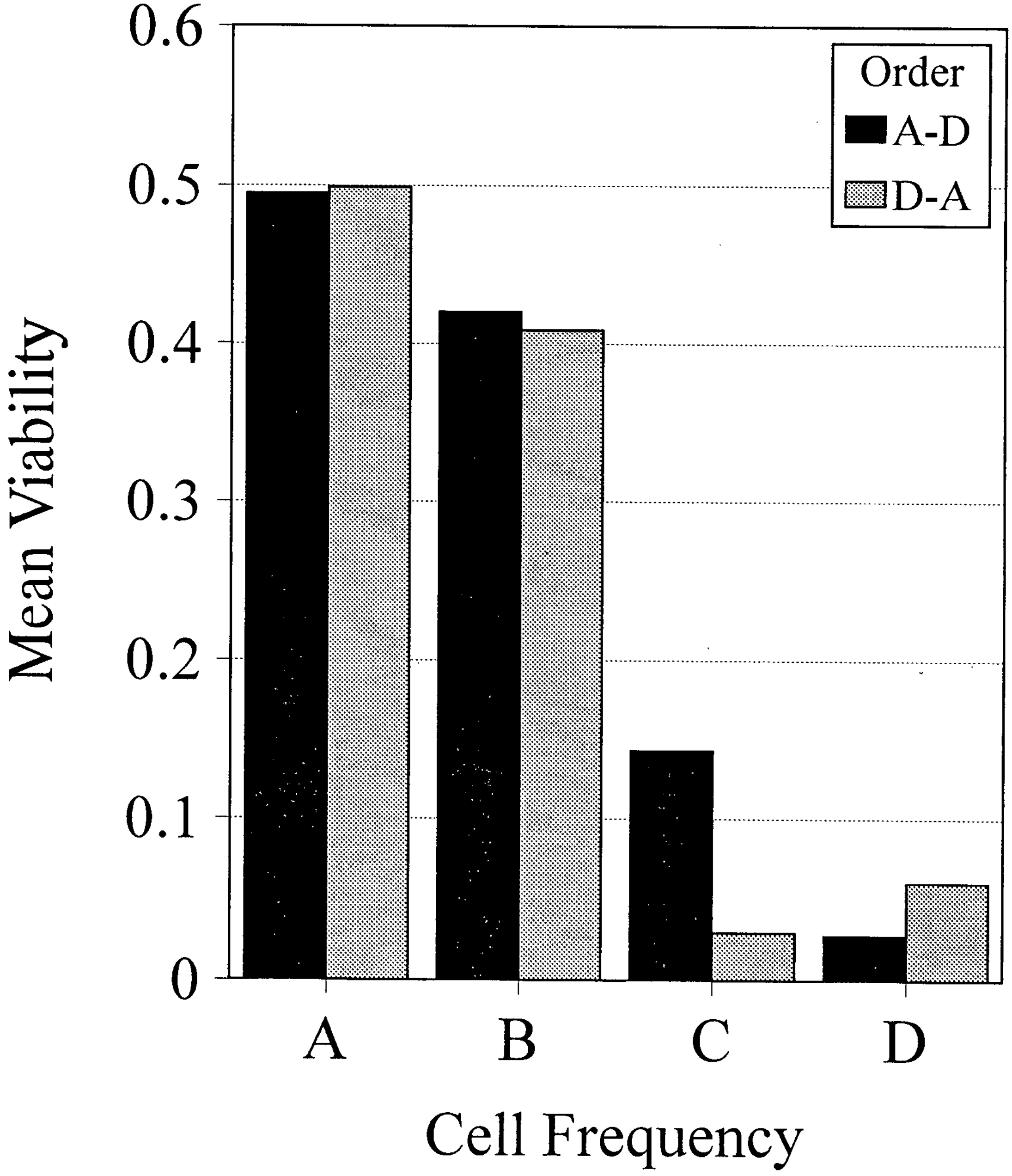

Figure 3. Mean viability of each cell frequency as a function of information presentation order in Experiment 3. (Note that the mean viabilities of *B* and *C* were multiplied by −1.)

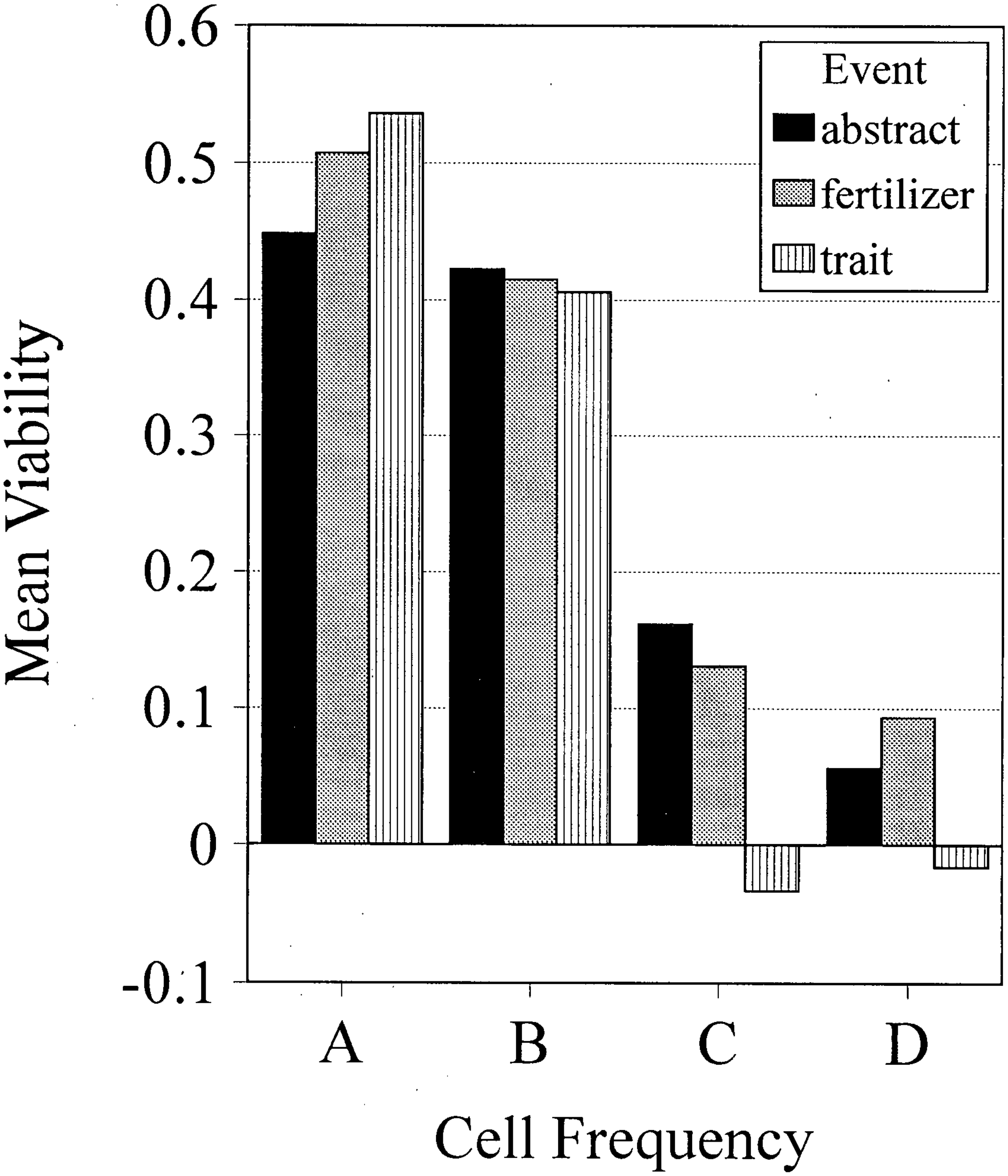

Figure 4. Mean viability of each cell frequency as a function of event type in Experiment 3. (Note that the mean viabilities of *B* and *C* were multiplied by −1).

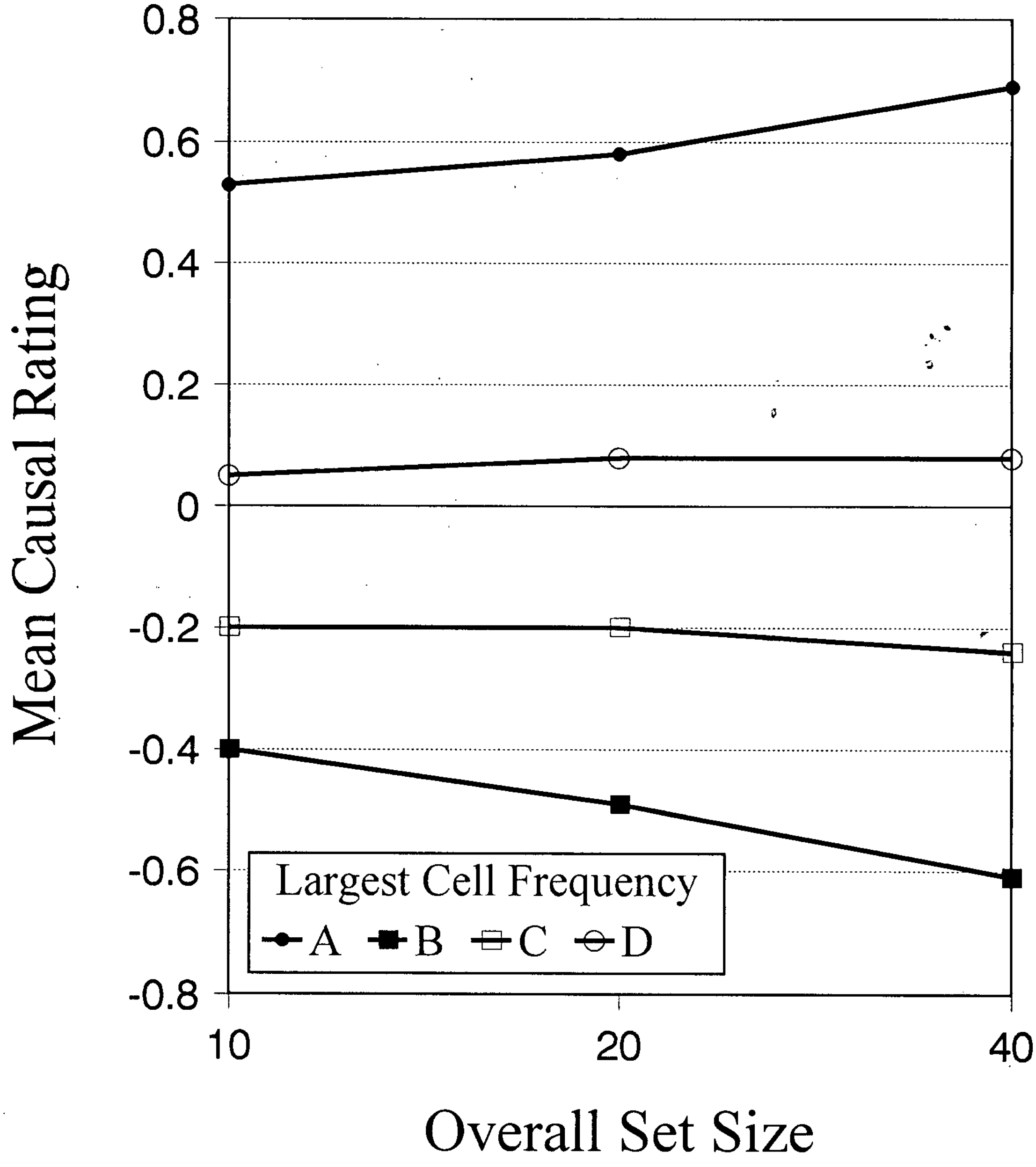

Figure 5. Mean causal ratings for stimulus subsets grouped by largest cell frequency as a function of overall set size in Experiment 3.

Facilitative Causality

	O_p	O_a	
I_p	**Cell A** +/+ Confirmation	**Cell B** +/– Sufficiency	+I, 3+:1–, Stest
I_a	**Cell C** –/+ Necessity	**Cell D** –/– Confirmation	–I, 1+:3–, Ntest
	+O, 3+:1–, Ntest	–O, 1+:3–, Stest	

Inhibitory Causality

	O_p	O_a	
I_p	**Cell A** +/+ Sufficiency	**Cell B** +/– Confirmation	+I, 3+:1–, Stest
I_a	**Cell C** –/+ Confirmation	**Cell D** –/– Necessity	–I, 1+:3–, Ntest
	+O, 3+:1–, Stest	–O, 1+:3–, Ntest	

Figure 6. Contingency tables depicting the relations between P-N, I-O, and S-N distinctions for facilitative and inhibitory causality. Columns and rows show the test type (P-N × I-O), the ratio of +:– cell features, and whether the contrast is an Stest or an Ntest.

APPENDIX A

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). Due to 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:
(a) Plants that were fertilized and that bloomed.
(b) Plants that were fertilized and that did *not* bloom.
(c) Plants that were *not* fertilized and that bloomed.
(d) Plants that were *not* fertilized and that did *not* bloom.

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-minute 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:
(a) People with the personality trait who started a conversation.
(b) People with the personality trait who did *not* start a conversation.
(c) People *without* the personality trait who started a conversation.
(d) People *without* the personality trait who did *not* start a conversation.