Memory & Cognition
2001, 29 (2), 193-208

How two causes are different from one: The use of (un)conditional information in Simpson's paradox

BARBARA A. SPELLMAN
University of Virginia, Charlottesville, Virginia

CHRISTY M. PRICE
University of Texas, Austin, Texas

and

JESSICA M. LOGAN
Washington University, St. Louis, Missouri

In a causally complex world, two (or more) factors may simultaneously be potential causes of an effect. To evaluate the causal efficacy of a factor, the alternative factors must be controlled for (or *conditionalized on*). Subjects judged the causal strength of two potential causes of an effect that covaried with each other, thereby setting up a Simpson's paradox—a situation in which causal judgments should vary widely depending on whether or not they are conditionalized on the alternative potential cause. In Experiments 1 (table format) and 2 (trial-by-trial format), the subjects did conditionalize their judgments for one causal factor on a known alternative cause. The subjects also demonstrated that they knew what information was needed to properly make causal judgments when two potential causes are available. In Experiment 3 (trial-by-trial), those subjects who were not told about the causal mechanism by which the alternative cause operated were less likely to conditionalize on it. However, the more a subject recognized the covariation between the alternative cause and the effect, the more the subject conditionalized on it. Such behavior may arise from the interaction between bottom-up and top-down processing.

Suppose that your child tests your baseball knowledge with the following question: If Player A has a higher batting average than Player B for the first half of the season *and* Player A has a higher batting average than Player B for the second half of the season, who has a higher average overall? If, like most people, you quickly and smugly answer "Player A, of course," you (along with 93% of our subjects) might be incorrect. As Table 1 illustrates, under some circumstances, it is in fact quite possible for Player B to have a higher average overall.

This example—that Player A does better in each half but Player B does better overall—is a demonstration of Simpson's paradox (after Simpson, 1951; see Pearl, 2000, for the history; example adapted from Paulos, 1988). Note that this "paradox" has two necessary components. Mathematically, it depends on the fact that the frequencies of at bats for the players differ in the two halves; if the frequencies were all equal, the problem could not arise (see Spellman, 1996b, for a proof). Conceptually, we see such examples as paradoxical when we are not sure whether to look at the parts (each half) or the whole (overall season) when making our judgments. For this baseball example, given all the statistics, one might want to claim that Player B is a better hitter because one cannot think of a reason that season half should be relevant to the analysis. But what if you were told that, in the second half of the season, a new kind of ball was introduced or that the size of the strike zone was changed? In these versions, because one might think that the type of ball or the size of the strike zone could affect performance, it might make more sense to claim that Player A is better because he is better under two different conditions that could be causally relevant to performance. (The Appendix further describes our baseball experiment.)

Simpson's paradox is not a mere conjuring trick by a clever experimenter; rather, it can arise unintentionally in laboratory data or even in "real-world" contexts. For example, Wainer (1986) reports that from 1980 to 1984, the mean SAT score of white test takers rose 8 points (from 924 to 932) and the mean SAT score of nonwhite test takers rose 15 points (from 802 to 817). What was the mean increase over all test takers? The statistically savvy reader

A portion of these results was presented at the 37th Annual Meeting of the Psychonomic Society, Chicago (1996). This research was supported by an NIMH FIRST Award and a Research Grant from the University of Texas to the first author. We thank Patricia Cheng, Denise Dellarosa Cummins, and Kelly Goedert for helpful comments on an earlier draft. We thank the Vietnamese restaurant on The Drag for many enjoyable B-4 lunches together. Correspondence concerning this article should be addressed to B. A. Spellman, Department of Psychology, University of Virginia, 102 Gilmer Hall, P. O. Box 400400, Charlottesville, VA 22904-4400 (e-mail: spellman@virginia.edu).

Table 1
Illustration of Simpson's Paradox in the Baseball Example

	Batting Average		
Player	First Half	Second Half	Overall Season Average
A	4 for 10 (**.400**)	25 for 100 (**.250**)	29 for 110 (.264)
B	30 for 100 (.300)	2 for 10 (.200)	32 for 110 (**.291**)

Note—Bold type indicates the higher average.

will guess something around 9 10 points, adjusting for the assumption that the number of white test takers is greater than the number of nonwhite test takers. However, the savviest reader will correctly say that there is not enough evidence to determine the answer. In fact, the answer is that the overall mean increased only 7 points (from 890 to 897). How can that be? From 1980 to 1984, the proportion of white test takers decreased and that of nonwhite test takers increased. As in the baseball example above, the unequal numbers allow for the paradox: The overall mean increase is less than the increase for each of the groups that make up the overall number.

A better known real-world example is that of admissions to the University of California, Berkeley. In the 1970s, Berkeley wanted to be sure it was not discriminating against female graduate school applicants. Despite its concern, overall, a far greater percentage of women than men were rejected; however, when admissions were looked at department by department, there seemed to be no discrepancy. The problem? Women tended to apply disproportionately to departments with high rejection rates (i.e., lots of applicants but few spots). (See Waldmann & Hagmayer, 1995, for a fuller description.) That describes the mathematical problem and solution; but the conceptual problem of whether the information should be considered at the department or the university-wide level remains. In contrast to the baseball example, in which it seems that we should aggregate and look at the overall season's average, here it seems that we should not aggregate and, instead, should look at the department level. For admissions, decisions are made at the department level, so department is a factor that is causally relevant to the outcome, whereas in our baseball example, season half is not causally relevant. Thus, rather than have a rule of whether to always (or never) aggregate, it seems that the right approach is to disaggregate only on causally relevant factors. (See Flexser, 1981, Hintzman, 1980, and Martin, 1981, for discussions relevant to the interpretation of contingency tables containing psychological data.)

SIMPSON'S PARADOX AND THE INTERPRETATION OF CAUSALITY

Simpson's paradox has serious implications for our understanding of the causes of events. For something to be a "cause," it is generally considered necessary, although not sufficient, for its presence to change the probability of the effect by some amount. Usually, when we talk about causes, we mean *facilitatory* causes—which raise the probability of the effect. For example, we say that smoking causes lung cancer (in part) because people who smoke have a greater probability of getting lung cancer than people who do not smoke. However, there are also *preventive* causes—for example, medications—which decrease the probability of an effect.

One way of characterizing the strength of a potential cause of an effect is by using contingencies; some people consider the ΔP contingency rule as the normative rule for computing causal strength (see Cheng, 1997, for a review and critique). Using the ΔP rule, one determines the probability of the effect given the presence of the proposed cause [$P(E|C)$] and the probability of the effect given the absence of the proposed cause [$P(E|{\sim}C)$]. The contingency is then computed as follows:

$$\Delta P = P(E|C) - P(E|{\sim}C).$$

The contingency is therefore bounded by -1 and 1.

As a numerical example, suppose that we wish to evaluate whether a particular blue liquid, advertised as a plant fertilizer, indeed causes plants to bloom. To do so, we determine whether the effect (blooming) is more probable when given the liquid than when not given the liquid. Suppose we have 40 identical plants. We pour the liquid on 20 of our 40 plants and find that 10 out of 20 of the treated plants bloom [$P(E|C) = .50$], but only 6 out of 20 of the untreated plants bloom [$P(E|{\sim}C) = .30$]. The contingency (ΔP) is therefore $.50 - .30 = .20$, and it seems that the liquid is a weakly effective fertilizer.

Suppose, however, that there are two potential causes of the plants' blooming. For example, perhaps we use a fertilizer, but some plants are in the sun and some are in shade. Should that potential alternative cause affect our ΔP evaluation of the causal efficacy of the blue fertilizer? In fact, some philosophers (e.g., Cartwright, 1979; Salmon, 1984) and psychologists (e.g., Cheng, 1993, 1997) argue that, in cases of multiple potential causes, ΔP is not normative and is the wrong rule to apply. They have suggested that when there are multiple potential causes of an effect, one should assess causality for each cause conditional on both the constant presence and constant absence of other potential causes (i.e., while controlling for those other potential causes). Obviously, it is not possible to ever know for certain that one has considered all alternative potential causes, but controlling for *known* alternative causes is a technique intentionally used by scientists to reduce the probability of errors in attribution. In fact, it seems that people (at least sometimes) know to do that in everyday attributions. For example, if you were to assert to a bunch of caffeine addicts that drinking coffee must cause lung cancer because people who drink lots of coffee get lung cancer more often than those who do not, the coffee drinkers would quickly point out that perhaps drinking coffee covaries with smoking, so it only looks like coffee causes lung cancer, whereas it is really smoking doing the causal work.

Mathematically, how is controlling for (or conditionalizing on) alternative causes done? To return to the plant

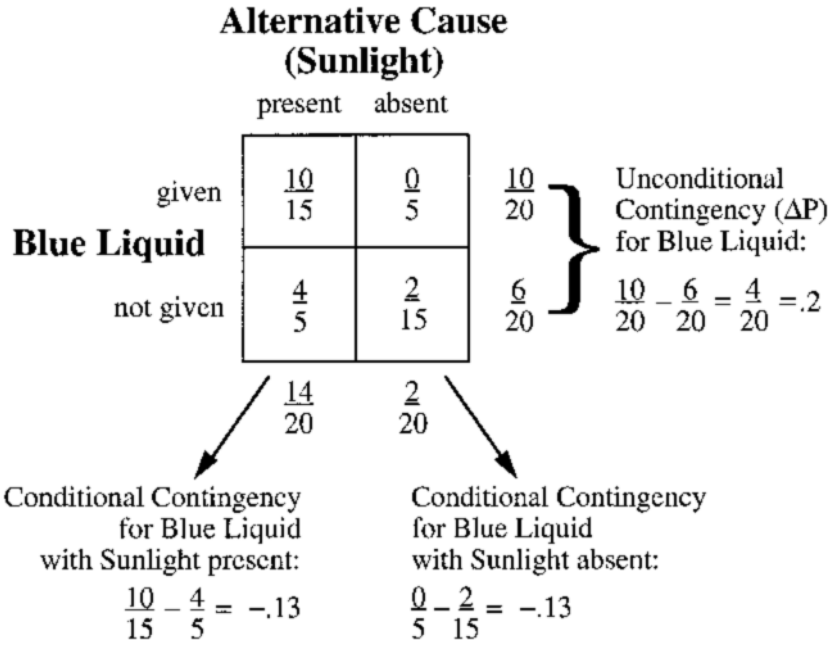


Figure 1. How to compute unconditional and conditional contingencies for the blue liquid in the example in the text. In this case, these contingencies are not equal, setting up the mathematical prerequisite for Simpson's paradox.

example from above, suppose that half our plants are in the sun and half are in the shade. We would like to evaluate the blue liquid regardless of location. According to the previous paragraph, to find the causal efficacy of the liquid, we should examine its effects both in the presence and in the absence of the alternative cause, sunlight. Consider the representation of frequencies shown in Figure 1. The fraction in each cell represents the number of times a plant blooms (numerator) over the number of times that combination of causes existed (denominator). The *unconditional* contingency (ΔP) for the liquid can be found by looking at the right-hand marginal and finding the difference between the proportion of times plants bloom when given the liquid (10/20) and the proportion of times plants bloom when not given the liquid (6/20)—therefore, as above, $\Delta P = .2$. However, suppose we wish to control for the sunlight and find the *conditional* contingencies for the liquid on the basis of the presence or absence of sunlight. When sunlight is present (left column), the difference between the proportion of times plants bloom when given the liquid (10/15 = .67) and the proportion of times plants bloom when not given the liquid (4/5 = .80) reveals a conditional contingency of −.13. When sunlight is absent (right column), the difference between the proportion of times plants bloom when given the liquid (0/5 = 0) and the proportion of times plants bloom when not given the liquid (2/15 = .13) again reveals a conditional contingency of −.13.[1] So, when we control for the alternative cause, the liquid looks like a preventive cause—it makes plants *less* likely to bloom. The marginals make it look as if the liquid is having an effect, because of the way in which *the application of the liquid covaries with the presence of the alternative cause* (i.e., because there are more events in the upper left and lower right cells). Once again, we see the possibility of Simpson's paradox, this time in a causal context: Does the blue liquid, in fact, cause plants to bloom?

Although mathematically, the paradox is evident (unconditional contingency = .20, conditional contingency = −.13), conceptually, this example does not feel much like a paradox; the alternative cause seems causally relevant, and we are not happy that our fertilized plants are blooming less in both sun and shade. Therefore, despite the positive unconditional contingency, we would ask for our money back on this so-called fertilizer.

Simpson's Paradox: The Mathematical Problem of Differing Base Rates

Would we expect people to be able to do this complex conditionalization when they are evaluating information?

Base rates of single events. Note that in order to recognize a Simpson's paradox, one has to be sensitive to base rates—the frequency of occurrence of events—when making statistical judgments. In the last 20 years, two different sets of results and perspectives seem relevant to the question of whether we would expect people to be sensitive to the frequency of occurrence of events. On the one hand, the literature on frequency judgments suggests that people are good at estimating the frequencies of the occurrence of events and that such information is automatically encoded (e.g., Hasher & Zacks, 1984). On the other hand, the *heuristics and biases* literature has doc-

umented many failures of people to use base rates in their reasoning (e.g., Tversky & Kahneman, 1974). Thus, it may seem that although frequency information is acquired, it is not often used. In his review of the base rate literature, Koehler (1996) discusses how this problem is (partially) resolved: Base rate neglect shows up most strongly when the information is embedded in a story involving statistics; however, when humans are asked to make judgments on the basis of presentation of trial-by-trial information (i.e., shown individual cases, rather than reading complete stories involving statistics), there is less (or no) base rate neglect (e.g., Gigerenzer, Hell, & Blank, 1988). Also (and relevant to the present studies), Koehler notes that subjects are more sensitive to base rates within causal than within noncausal cover stories (Spellman, 1996c; Tversky & Kahneman, 1982).

Two known potential causes. Of course, problems involving Simpson's paradox involve more than just keeping track of the frequencies of single events; what is also important is keeping track of the covariation between each of the causes and the effect and, possibly, the covariation of the causes with each other.

People have demonstrated the ability to use such covariation information and to conditionalize on alternative causes. For example, Spellman (1996a) asked subjects to view trial-by-trial presentations of neither, one, or both of two fertilizers (blue and red) being poured onto plants. The subjects' task was to judge how effective the fertilizers were. In the three conditions of Experiment 1, the blue liquid's unconditional contingency was 0, but the conditional contingency varied downward across conditions: Condition 1, .33; Condition 2, 0; and Condition 3, −.33. In accordance with the predictions of a conditional contingency analysis, subjects' causal ratings for the blue liquid decreased across those conditions. In the three conditions of Experiment 2, the unconditional contingency for the red liquid was different in all conditions—Condition 4, .5; Condition 5, 0; and Condition 6, −.5—but its conditional contingency was 0 in all conditions. In this set-up, causal ratings for the red liquid did *not* vary across conditions, again in accordance with the predictions of a conditional contingency analysis.

In fact, the data from several articles that claimed to show that people were poor reasoners because they deviated far from ΔP when reasoning about multiple causes of effects (e.g., Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Chapman, 1991; Chapman & Robbins, 1990; Price & Yates, 1993, 1995) can be reanalyzed to show that people seem to be using the "smarter" conditional contingency strategy (Cheng, 1993; Melz, Cheng, Holyoak, & Waldmann, 1993; Shanks, 1993, 1995; Spellman, 1993, 1996a, 1996b).

All of the above involve cases in which the causes are known. However, obviously, in a world filled with an infinity of potential causes and causal combinations, it is not possible to keep track of all that information for all possible potential causes. So, when is conditionalizationlikely to occur?

Simpson's Paradox: The Conceptual Problem of Whether Conditionalization Is Appropriate

Two different lines of research address the question of when we conditionalize and what we conditionalize on.

Schaller and colleagues (Schaller, 1992a, 1992b; Schaller & O'Brien, 1992) have investigated the use of something akin to conditionalization—what he calls "intuitive analysis of covariance"—in tasks that do not involve causal reasoning. For instance, in a study that is reminiscent of our baseball example, subjects were presented with information about the racquetball prowess of two potential doubles partners. Player 1 performs better than Player 2 in both League A (1 wins 20/80, 2 wins 0/20) and League B (1 wins 20/20, 2 wins 60/80). However, Player 2 shows better overall performance (total wins of 40/100 vs. 60/100). The subjects were asked to rate which player was better; the answer depended on whether league was taken into account. In studies like this one, Schaller has shown that whether subjects covary out a factor when making ratings may depend on motivation, sample size, perceived relevance of the alternative factor, instructions, and time available to process information.

Waldmann and Hagmayer (1995) used a causal cover story to examine conditionalization. Subjects had a list of 80 pieces of information about the potential causes of an effect in front of them (e.g., type of plant, whether it was watered, whether it grew). When rating whether watering helped the plants to grow, the subjects were more likely to take type of plant into account when (1) given a hint by the experimenter that type of plant might matter or (2) the information was grouped by plant type (thus making either the factor or the pattern of results more obvious to the subjects).

THE PRESENT RESEARCH

General Method

The present experiments add several new pieces to our understanding of Simpson's paradox. Because all three experiments used similar cover stories and contingencies, we will first describe the similarities before describing the particulars of each experiment.

The experiments involved causal cover stories, in which people might need to judge the causal efficacy of one cause conditional on the presence or absence of another cause. In these experiments, subjects learned about two potential causes of plants' blooming: (1) treatment with a blue liquid that might be a fertilizer and (2) being planted in a pot that has a star emblem on it. Exactly how the star emblem could affect the plant was not obvious to the subjects until explained; that is, the subjects had no preexisting knowledge or theory about how the emblem could affect the plants. In these experiments, some subjects were told that the emblems are part of a mechanism that extends inside the flowerpot to inject the soil with a fine mist that might be a fertilizer, whereas other subjects were told nothing about the emblems. The subjects received the following information about each of 80 plants: whether it is treated with the liquid, whether it is in a star pot, and whether it blooms. The subjects then made several kinds of judgments, including causal efficacy ratings for how the liquid and the emblem affect blooming and confidence ratings on those efficacy judgments.

In Experiment 1, we examined whether subjects conditionalize their causal efficacy judgments when the information is provided in

a summary table format. Such a format allowed us to present subjects either with all of the conditionalized information (as is usually done) or with only marginal information (a new method). The full-table presentation allowed us to examine whether subjects conditionalize; the marginals-only presentation allowed us to examine whether subjects implicitly or explicitly know that they need the conditional information in order to make good causal judgments.

Experiments 2 and 3 switched to trial-by-trial presentations of information. In Experiment 2, we evaluated whether people conditionalize the efficacy of the liquid on the star emblem—an odd, but plausible, alternative cause. In Experiment 3, we looked at the same information presented under different cover stories—either explaining the star emblem mechanism or not. We asked whether a top-down theory of how the star emblem might be causally related to blooming is necessary for conditionalization or whether subjects who had no top-down theory might still become aware of the effect of the star emblem and then conditionalize their contingency judgments for the blue liquid on it.

In our experiments, we had a *baseline*, or *equal*, condition, in which some subjects learned information with the same unconditional contingencies and the same marginal information as in the experimental (i.e., Simpson's paradox) condition. In the equal condition, the frequency of events in the four cells was equal, and the unconditional and conditional contingencies were equal, so there could not be a paradox. This condition tells us what subjects do with the information when there is no need to conditionalize. We can compare the equal condition to what subjects do when conditionalization is a possibility, to see how much conditionalizing is occurring. In addition, in all the experiments, the subjects made judgments on a rating scale that went from −100 (*strong negative cause*) to 0 (*noncausal*) to +100 (*strong positive cause*), thus paralleling the −1 to 0 to 1 contingency measure. Unlike previous research, which often used relative ratings, we can use this scale to (1) compare causal ratings with the suggested "normative" standards (i.e., unconditional and conditional ΔP) and (2) compare causal ratings across experiments. We also had subjects make confidence judgments on their causal ratings, to discover whether they were sensitive to the quality of the information they were getting.

In Experiments 2 and 3, the subjects made trial-by-trial predictions of whether they thought the plant would bloom before getting feedback about whether it actually did. This technique, analogous to that used in many category-learning experiments, may lead to better encoding of the information than merely watching the presentation of covariation information. The two trial-by-trial experiments also had an additional dependent measure: Subjects made predictions about how the potential causes would affect future plant blooming. From these predictions we can get a non-rating-scale measure of their subjective *derived contingency*.

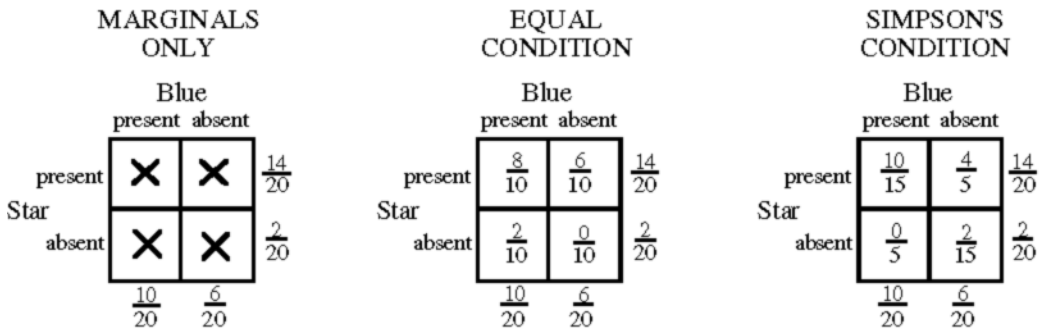
EXPERIMENT 1
The Importance of Covariation Information: Do Subjects Know They Need to Conditionalize?

Experiment 1 was designed to address the following four questions, using a summary table format. Will subjects use unconditional or conditional contingencies in judging the efficacy of one potential cause (the blue liquid) when given an odd, but plausible, alternative causal mechanism? Will subjects' judgments, on a −100 to 100 scale (where 0 is *noncausal*) reflect the normative contingencies? Will subjects' confidence in their judgments differ depending on whether (1) the unconditional and conditional contingencies are equal and (2) whether they have information about how the causes covary? Do subjects know that they need information about how causes covary with each other before they can assess the causal efficacy of each?

To determine whether subjects would conditionalize, all the subjects read the same cover story about the blue liquid and star emblem possibly acting as fertilizers. They saw a table describing the relation between the causes and the effect. Mathematically, what remained constant across conditions were the marginal totals in the tables and, therefore, the unconditional contingency for the liquid.[2] (The unconditional contingency for the liquid was always .20; see Figure 2.) What varied between subjects was the information in the cells of the tables and, therefore, the conditional contingency for the liquid. In the equal condition, the liquid was clearly a fertilizer; both the unconditional and the conditional contingencies were .20. In the Simpson's paradox condition, however, whether the liquid should be viewed as a fertilizer depends on whether or not its efficacy is conditionalized on the star emblem. Overall (i.e., according to the marginals), it seems that the liquid helps blooming and that $\Delta P = .20$ (as in the equal condition). But when the efficacy of the liquid is conditionalized on the presence or absence of the star emblem, it can be seen that the liquid actually decreases the probability of blooming and that the conditional $\Delta P = -.13$. (The same information was used in the example presented in the text.)

To address the question of whether subjects know that they need information about how causes covary with each other, we created a third condition, called the *marginals-only* condition. In the equal and Simpson's paradox conditions, the subjects received information about how often the plant bloomed given each combination of treatments. For example, in the equal condition, they were told that of the 80 plants, 16 of the 20 that were in the star pots and got blue liquid bloomed, 4 of the 20 that were in plain pots and got blue liquid bloomed, and so forth. However, in the marginals-only condition, the subjects were given information only about one treatment at a time and were not given the combination information. For example, they were told that 20 of the 40 plants that got blue liquid bloomed (some were in star pots, and others were not), 28 of the 40 plants that were in the star pots bloomed (some got blue liquid, whereas others did not), and so forth. Thus, they do not have the information about whether the causes covary with each other.

By not providing cell information, we could find out various things. First, we could examine the causal ratings that the subjects would make when they were lacking the cell information. Without that information, the conditional contingency could not be determined. In the baseball example, it seemed as if the subjects just assumed that the cell frequencies (i.e., the base rates) were equal. Would subjects do so with this unfamiliar cover story? If they were to assume equal frequencies, the unconditional and conditional contingencies would be equal (and iden-



Contingencies

	BLUE	BLUE	BLUE
UC	.20	.20	.20
CC	?	.20	−.13
	STAR	**STAR**	**STAR**
UC	.60	.60	.60
CC	?	.60	.67

Figure 2. The cell information and the unconditional and conditional contingencies for the marginals-only, equal, and Simpson's paradox conditions used in all of the experiments.

tical to those in the equal condition). Thus, we would expect causal ratings in the marginal-only condition to be similar to those in the equal condition.

Second, we could determine whether subjects would know that they needed the cell information to make an accurate judgment. That knowledge could reveal itself in either (or both) of two ways. One would be in the confidence ratings that the subjects made about their efficacy judgments. If the subjects were less confident in the marginals-only condition than in the other two conditions, that would reflect that, at least *implicitly*, subjects knew that they needed more data to make the efficacy judgments. The second would be in their answers to a question following the confidence judgment that asked whether there was anything they would like to know that would make them more confident in their judgments. If the subjects were to state that they wanted information about how the causes act in combination, that would reflect *explicit* knowledge that they needed more data to make the efficacy judgments.

Method

Subjects

The subjects were 66 University of Texas undergraduates who participated in partial fulfillment of an introductory psychology course requirement. The subjects were tested in small groups of varying sizes. Each subject received a booklet containing this experiment along with other short reasoning tasks. They were encouraged to take as much time as they needed to answer all questions.

Design

The subjects were randomly assigned to one of three conditions: equal ($n = 24$), Simpson's paradox ($n = 21$), and marginals-only ($n = 21$). In all the conditions, the unconditional contingency for the blue liquid was .20. In the equal condition, the conditional contingency for the blue liquid was also .20; in the Simpson's paradox condition, the conditional contingency was −.13; in the marginals-only condition, the conditional contingency could not be determined (see Figure 2).

The contingencies for the star emblem remained fairly equal across conditions. In all the conditions, the unconditional contingency for the star emblem was .60. In the equal condition, its conditional contingency was also .60; in the Simpson's paradox condition, its conditional contingency was .67; in the marginals-only condition, its conditional contingency could not be determined (see Figure 2).

Materials

The subjects read the following story:

> Imagine that while looking though the garage of the house you have just rented, you find some very interesting-looking containers of liquid. Your landlady tells you that some of them are very expensive plant-treatment liquids and some of them are just colored water. Of the plant-treatment liquids, she remembers that some of them are flower-growth stimulators (fertilizers) and some are flower-growth inhibitors (weed killers) and that the liquids came in various strengths—but she does not remember which liquid is which. She does want you to find out, however, and is willing to reduce your rent if you can figure it out.
>
> The landlady has a bunch of plants in her greenhouse on which she has poured the BLUE liquid. However, you should note that some of the plants are in special STAR pots. STAR pots have a built-in mechanism

(in the shape of a star) that automatically releases chemicals to the plants at regular intervals. Here is what happened to her plants.

The subjects were then given a summary table of the treatments and outcome (blooming), with information about 80 plants (i.e., double the information in Figure 2).

Equal condition. The subjects learned about blooming in all possible treatment combinations:

20 were in the STAR pots and got BLUE liquid. 16 bloomed.

20 were in plain pots and got BLUE liquid. 4 bloomed.

20 were in STAR pots and did not get blue liquid. 12 bloomed.

20 were in plain pots and did not get blue liquid. 0 bloomed.

Simpson's paradox condition. The subjects learned about blooming in all possible treatment combinations:

30 were in the STAR pots and got BLUE liquid. 20 bloomed.

10 were in plain pots and got BLUE liquid. 0 bloomed.

10 were in STAR pots and did not get blue liquid. 8 bloomed.

30 were in plain pots and did not get blue liquid. 4 bloomed.

Marginals-only condition. The subjects learned about blooming for each potential cause separately and were not given information about treatment combinations:

40 got BLUE liquid (some were in STAR pots and others were not). 20 bloomed.

40 did not get blue liquid (some were in STAR pots and others were not). 12 bloomed.

---

40 were in the STAR pots (some got BLUE liquid while others did not). 28 bloomed.

40 were in plain pots (some got BLUE liquid while others did not). 4 bloomed.

After reading the story and the outcome, the subjects made three types of judgments. First, they rated the effectiveness of the liquid and the emblem on a scale from −100 to 100, in which negative numbers indicated that the liquid was a flower growth inhibitor (making the plant *less* likely to bloom), zero meant that it had no effect (colored water), and positive numbers meant that it was a flower growth stimulator (making the plant *more* likely to bloom). Then, they made confidence ratings on their efficacy judgments for the liquid and emblem from 0 (*no confidence*) to 10 (*total confidence*). Finally, they were asked if there was anything they would like to know that would make them more confident in their judgments.

Results and Discussion

Three subjects, all in the equal condition, used the rating scales incorrectly and were excluded from the analyses.

Efficacy Ratings

For the blue liquid, the subjects' mean ratings differed across conditions, as is shown in Table 2. The first question is whether subjects conditionalized their ratings of the blue liquid on the star emblem when the conditionalizing information was available. They did. The mean rating for the blue liquid in the equal condition ($M = 29.9$) was positive and was significantly higher than the negative mean rating in the Simpson's paradox condition [$M = -8.6$; $t(40) = 2.98$, $p < .005$]. In the Simpson's paradox condition, the rating was also significantly lower than the actual value for its unconditional contingency [times 100; $t(20) = 2.45$, $p < .05$], but did not significantly differ from the actual value for its conditional contingency [times 100; $t(20) = 1.79$, $p > .08$]. In the equal condition, the rating did not differ from its actual un/conditional contingency [$t(20) < 1$].

The second question is how the subjects gave ratings when they had only the marginal information. In this condition, the mean causal strength rating ($M = 21.9$) did not differ from that for the equal condition [$t(40) < 1$]. Nor did the ratings in the marginal condition differ from .20—the unconditional contingency value for this condition [$t(20) < 1$]. The Simpson's paradox condition demonstrates that, with this cover story, subjects will use conditional rather than unconditional frequencies. What the mean rating of 21.9 suggests is that in the absence of cell information, subjects make the assumption that the cells all have equal frequency of occurrence. Given just the marginal information, plus the assumption that all the cells have equal frequencies, subjects can extract conditional contingencies—which, under such a procedure, would be equal to the unconditional contingency.[3]

The subjects' ratings for the star emblem were very similar across conditions. Mean causal strength ratings did not significantly differ across the equal ($M = 65.0$), Simpson's paradox ($M = 69.5$), and marginals-only ($M = 50.7$) conditions, as was indicated by an omnibus F test [$F(2,60) = 1.32$, n.s.]. Because the unconditional and conditional contingencies for the star emblem were kept (nearly) constant across conditions, no difference in ratings was predicted.[4] None of these ratings was significantly different from either their unconditional or their conditional (when determinable) contingency (all ps $> .20$).

Thus, the subjects appeared to be conditionalizing on alternative causes, and their rating scale data were quite close to the actual conditional contingencies.

Confidence Ratings and Information Requested

The confidence ratings were designed to get at two questions about the relation between information and confidence.

First, we found that confidence was lower when there was a discrepancy between the unconditional and the conditional contingencies. When comparing the equal and the Simpson's paradox conditions for the blue liquid, the subjects were significantly less confident in the Simpson's paradox condition ($M = 5.6$), in which there was a big discrepancy between the unconditional and the conditional contingencies, than in the equal condition ($M = 7.2$), in which those contingencies were equal [$t(40) = 2.14$, $p < .05$]. For the star emblem, there was no difference in the subjects' confidence between the Simpson's paradox ($M = 7.5$) and the equal ($M = 7.4$) conditions [$t(40) < 1$]. Note that for the star emblem, unlike for the blue liquid, the unconditional and the conditional contingencies were nearly equal. Thus, although the subjects in the Simpson's paradox condition did use the conditional contingency when

Table 2
Contingencies and Results From Experiments 1–3

	Condition				
	Marginals Only	Equal	Simpson's Paradox (Tell)	Simpson's Paradox (Look)	Simpson's Paradox (No Information)
			Blue Liquid		
Contingencies					
Unconditional	.20	.20	.20	.20	.20
Conditional	?	.20	−.13	−.13	−.13
Experiment 1 (table)					
Ratings					
Mean	21.9*	29.9*	−8.6†		
SE	10.2	5.5	11.7		
Confidence	5.5*	7.2†	5.6*		
Experiment 2 (trials)					
Ratings					
Mean		20.6*	−8.0†		
SE		4.8	5.6		
Confidence		6.5	6.6		
Derived contingency					
Mean		11.7*	−2.8†		
SE		4.2	4.3		
Experiment 3 (trials)					
Ratings					
Mean			−12.5*	8.7†	22.4†
SE			5.1	3.8	6.0
Confidence			6.6	6.2	5.7
Derived contingency					
Mean			−5.5*	4.3*†	12.9†
SE			4.0	2.5	5.2
			Star Emblem		
Contingencies					
Unconditional	.60	.60	.60	.60	.60
Conditional	?	.60	.67	.67	.67
Experiment 1 (table)					
Ratings					
Mean	50.7	65.0	69.5		
SE	10.6	4.0	9.5		
Confidence	5.3*	7.4†	7.5†		
Experiment 2 (trials)					
Ratings					
Mean		61.6	57.8		
SE		7.4	8.6		
Confidence		7.0	7.1		
Derived contingency					
Mean		42.7	43.9		
SE		5.9	6.1		
Experiment 3 (trials)					
Ratings					
Mean			65.0*	47.9*†	37.4†
SE			6.5	6.2	8.0
Confidence			7.5*	5.7†	5.8†
Derived contingency					
Mean			53.7*	45.4*	15.9†
SE			5.1	5.6	6.1

Note—Different superscripts in a row denote significantly different results.

there was a discrepancy, they were less confident in those judgments than they were when there was no such discrepancy.

Second, we found that the subjects in the marginal-only condition were less confident of their judgments than were the subjects in the equal condition. The confidence judgments were an implicit measure of that belief, and the information question was an explicit measure of that belief. For both the blue liquid ($M = 5.5$) and the star emblem ($M = 5.3$), the subjects were significantly less confident in the marginals-only condition than in the equal condition [$t(40) = 2.74$, $p < .01$, and $t(40) = 3.0$, $p < .005$, respec-

tively]. In addition, 33% of the subjects in the marginals-only condition requested cell information in their comments.

Therefore, many subjects seem to be aware of the necessity for information about how the causes covary with each other in order to make a good causal judgment. For both causes, the subjects were less confident when given only marginal information, indicating that implicitly they were aware that the quality of the information was poor. The request for information indicates that at least one third of the subjects were explicitly aware of that problem.

EXPERIMENT 2
Simpson's Paradox With Trial-by-Trial Information

In Experiment 1, the subjects used conditional contingencies when evaluating information presented in a summary table format. Experiment 2 used the same cover stories and contingencies as Experiment 1 but presented the information in a trial-by-trial manner. We eliminated the marginals-only condition and added a new judgment, called *derived contingencies*.

There are reasons to believe that subjects might not be as good at the trial-by-trial version of the task as on the summary table version of the task (e.g., Kao & Wasserman, 1993; Ward & Jenkins, 1965). One suggestion for why that might be is that trial-by-trial presentations place a higher memory load on subjects (Shaklee & Mims, 1982). That assumes, however, that the subject is using a strategy of counting and remembering different kinds of trials. Note two differences between those experiments and the present one: (1) In the previous experiments, the subjects judged the relation between one cause and one effect; here, they judge the relation between two causes and one effect; and (2) in the previous experiments, the subjects were passive viewers of the stimuli; here, they must make a prediction on each trial. Whereas the former difference suggests that our task should be even more difficult than trial-by-trial covariation tasks in previous experiments, the latter difference suggests that subjects might encode the covariation information better in our task (Koehler, 1996).

We were interested in many of the same questions as in Experiment 1, this time for a trial-by-trial presentation: Would the subjects use unconditional or conditional contingencies in judging the efficacy of the liquid fertilizer when given an odd, but plausible, alternative causal mechanism? Would their judgments on a −100 to 100 scale (where 0 is *noncausal*) reflect the contingencies? Would the subjects' confidence in their judgments differ depending on whether the unconditional and the conditional contingencies were equal?

We were also interested in whether subjects' knowledge of the causal efficacy of the blue liquid and star emblems would be reflected in a dependent measure that should more closely reflect the subjects' frequency knowledge and contingency beliefs than do the rating scale (which is not a direct measure of that knowledge). As was pointed out by Price and Yates (1993), much research on contingency judgments asks the subject to rate on a scale either from 0 to 100 or from −100 to 100 the effect of (or predictiveness of) some variable on another variable. Such ratings need not have any particular correspondence with the actual contingencies used in the experiments. Therefore, we decided to ask subjects to make judgments on the basis of frequency information and to compute their derived contingencies from those judgments. For example, for the blue liquid, after making the efficacy rating, the subjects were told to imagine that the liquid was poured on another 100 plants and were asked how many plants out of that 100 would bloom. They were then told to imagine that the liquid was *not* poured on another 100 plants and were asked how many plants out of that 100 would bloom.[5] The derived contingency is the difference between the subjects' responses to the liquid-present and the liquid-absent questions.

Method

Subjects

The subjects were 45 University of Texas undergraduates who participated in partial fulfillment of an introductory psychology course requirement.

Design

The subjects were randomly assigned to one of two conditions: equal or Simpson's paradox. The contingencies for these conditions were the same as those in Experiment 1 (see Figure 2).

Materials and Procedure

The subjects were run individually on computers.

Cover story. The cover story was identical to that in Experiment 1. After reading the story, the subjects were told that they would be using a scale to rate the blue liquid's and star emblem's effects on the plants; these scales were explained in detail in the instructions.

Trial-by-trial presentations. In each trial, the subjects saw a picture of a plant in either a pot decorated with a star emblem or a plain undecorated pot, with the liquid either being poured on the plant or not being poured on the plant. At the bottom of the screen was the question, "Do you think the plant will produce a flower?" The subjects were given as much time as they needed to respond with a *yes* or *no* keypress. The subjects were then given feedback for 5 sec on the next screen, where the same plant was shown either blooming or not blooming. A blocked randomization technique was used for presenting the trial-by-trial information. Each of five blocks consisted of 20% of the information from each cell given in Figure 2, with order of presentation randomized within each block, yielding a total of 40 trials across blocks. The same blocks, but with new random presentation orders, were presented again after the initial ratings, for a total of 80 trials per experiment.

Efficacy and confidence ratings. After the initial 40 trials, the subjects were asked to make a practice rating judging the efficacy of the blue liquid and then give a confidence rating for the judgment. They rated the blue liquid's efficacy on a scale from −100 to 100, in which negative numbers indicated that the liquid was a flower growth inhibitor (making the plant *less* likely to bloom), zero meant that it had no effect (colored water), and positive numbers meant that it was a flower growth stimulator (making the plant *more* likely

to bloom). The confidence rating was made on a scale from 0 (*no confidence*) to 10 (*complete confidence*) that the judgment was correct. After another set of 40 trials, the subjects were asked to give an actual rating of the efficacy of the blue liquid and to give their confidence rating again.

Derived contingencies. After the final rating of the efficacy of the blue liquid, the subjects were told to imagine that the blue liquid was poured on another 100 plants and were asked how many plants out of that 100 would bloom. They were then asked to suppose that the blue liquid was *not* poured on another 100 plants and were asked how many plants out of that 100 would bloom. The subjects were told to assume for both predictions that the pots had the same ratio of pots with star emblems to undecorated pots as in the trial-by-trial presentations. Subtracting the subject's response for the liquid-absent situation from their response for the liquid-present situation yields a derived contingency.

Judgments of the star emblem. After making the final judgments for the blue liquid, the subjects were asked the analogous questions about the star emblem. First, they judged the efficacy of the star emblem and rated their confidence in that judgment. Then, subjects were asked derived contingency questions with respect to 100 pots with and 100 pots without the star emblem, assuming the same ratio of liquid-given to no-liquid-given as in the trial-by-trial presentations.

Results and Discussion

Data from 5 subjects, 2 from the equal condition and 3 from the Simpson's paradox condition, were excluded from the analyses because the subjects typed in numbers outside the boundaries for the ratings scales. In each condition, the judgments after 80 trials from the remaining 20 subjects were analyzed.

Efficacy Judgments

The subjects' ratings for the blue liquid were significantly different in the equal and Simpson's paradox conditions. In the equal condition, the mean was positive ($M = 20.6$), whereas in the Simpson's paradox condition, the mean was negative [$M = -8.0$; $t(38) = 3.86$, $p < .001$]. Each of the means is quite close to (and not significantly different from) its actual conditional contingency, as is shown in Table 2 [$t(19) < 1$] for each. Thus, it seems that when judging the blue liquid, the subjects used conditional rather than unconditional contingency information in their ratings. They were willing to conditionalize on the star emblem, a feature that might not be perceived as causally relevant to plants blooming, when a causal mechanism was provided.

The subjects' ratings for the star emblem in the equal ($M = 61.7$) and Simpson's paradox ($M = 57.8$) conditions were not significantly different [$t(38) < 1$]. As in Experiment 1, because the unconditional and conditional contingencies for the star emblem were kept (nearly) constant across conditions, no difference in ratings was predicted. Again, neither of these ratings was significantly different from either its unconditional or its conditional contingency (all ps $> .20$).

Considering the equal and Simpson's paradox conditions, Experiment 2's trial-by-trial format thus replicates Experiment 1's summary table format in that subjects will conditionalize their causality judgments of one cause on another. As can be seen in Table 2, the ratings for the blue liquid were very similar across the two experiments, and all ratings were quite close to the appropriate conditional contingency value.

Confidence Ratings

The subjects' confidence ratings did not vary across conditions for either the blue liquid or the star emblem [both $t(38) < 1$]. Thus, despite the discrepancy between the unconditional and the conditional contingencies for the blue liquid in the Simpson's paradox condition, the subjects were just as confident of their judgments as they were for the blue liquid in the equal condition, where there was no such discrepancy. This result is different from that in Experiment 1, in which the subjects were less confident in judging the blue liquid in the Simpson's paradox condition.

Derived Contingencies

The contingencies derived from the subjects' responses followed the same pattern as the subjects' ratings, although they were smaller in magnitude. For the blue liquid, the derived contingency was significantly higher in the equal ($M = 11.7$) than in the Simpson's paradox ($M = -2.8$) condition [$t(38) = 2.41$, $p < .05$]. For the star emblem, derived contingencies did not differ between the equal ($M = 42.7$) and the Simpson's paradox ($M = 43.9$) conditions [$t(38) < 1$]. It appears that these derived contingencies mimic the actual conditional contingencies and the subjects' ratings, although the derived contingencies are less extreme. In fact, the ratings and derived contingencies were significantly correlated [liquid, $r(40) = .59$, $p < .0001$; emblem, $r(40) = .33$, $p < .05$]. We do not know what effect making causal ratings first had on the later derived contingency judgments. Nevertheless, we see that the subjects conditionalized not only when using a rating scale to judge the past, but also when making actual frequency predictions of the future.

EXPERIMENT 3
Top-Down Versus Bottom-Up Reasons for Conditionalization

The previous experiments demonstrate that subjects will conditionalize their judgments of the blue liquid (a typical cause of blooming) on the presence or absence of the star emblem (an odd but plausible cause of blooming) when provided with an explanation of how the star emblem could work. In Experiment 3, we withheld the explanation of the star emblem's mechanism from some of the subjects (the *no-information* condition). The question was whether subjects without the mechanism information would conditionalize their judgments on the emblem (an implausible cause of blooming).

There were several possible outcomes. One possibility was that subjects in the no-information condition would not conditionalize at all, because a causal mechanism is necessary for conditionalization. Proponents of the mechanism view of causal reasoning have argued that knowledge of an underlying causal mechanism—that is, an understanding of *how* a cause might produce an effect—is necessary for

people to infer a causal relation from a covariation (Ahn, Kalish, Medin, & Gelman, 1995; Bullock, Gelman, & Baillargeon, 1982). If subjects in the no-information condition know of no causal mechanism to explain the covariation between the emblem and blooming, they should not consider that relation as causal. Without such a reason to believe that the emblem might be relevant to blooming, subjects will see only the unconditional contingency between the liquid and the blooming (and so will behave like the subjects in the equal condition in Experiment 2).

A second possibility was that subjects in the no-information condition would fully conditionalize on the star emblem even without any causal theory (and so would behave like the subjects in the Simpson's paradox condition of Experiment 2). This possibility was unlikely, given the results of Waldmann and Hagmayer (1995), which showed that providing subjects with a causal theory increased conditionalization,[6] and of White (1995), which showed that prior beliefs about the causal power of a factor can affect the interpretation of covariation information.[7]

A third possibility was that, during the course of the experiment, (at least some) subjects might learn purely bottom-up that the emblem was causally related to blooming and then conditionalize on it. We could assess that learning in several ways. First, we could determine whether the subjects in the no-information condition rated the emblem as having any effect at all; giving the emblem an efficacy rating above zero would indicate learning of the covariation. Second, we could see whether those subjects then conditionalized their efficacy judgments of the liquid on the emblem. Third, we could explicitly ask the subjects whether they noticed any relation between the emblem and blooming.

Method

Subjects

The subjects were 62 University of Texas undergraduates who participated in partial fulfillment of an introductory psychology course requirement.

Design

The subjects were randomly assigned to one of three conditions, which differed only in how much information was given to the subjects about the star emblem: tell, look, or no-information. All the conditions used the same contingencies— those from the Simpson's paradox conditions of the previous experiments. The tell condition's cover story was identical to the cover story of Experiment 2: The subjects were told about the star emblem's mechanism and that they would have to figure out how both the blue liquid and the star emblem affected the plants. The subjects in the tell condition were informed that they would use the rating scales to indicate the efficacy of both the liquid and the emblem. In the look condition, the subjects were simply told that some pots were decorated with the emblem and some were not. In the no-information condition, the subjects were told nothing about the emblem; it was not mentioned at all in the instructions. In these latter two conditions, the subjects were instructed that they would be using the rating scale for the liquid, but rating the emblem was not mentioned.

Materials and Procedure

The materials and procedures were identical to those in Experiment 2.

Cover stories. In all the conditions (tell, look, and no-information), the subjects were given the same general instructions and cover story as those in Experiment 2. As described above, the conditions differed only in the amount of information given to the subjects about the star emblem.

Efficacy and confidence ratings. The efficacy and confidence ratings were identical to those in Experiment 2.

Free response booklet. After finishing the computer portion of the experiment, the subjects filled out a short questionnaire that asked them what strategy they had used to decide whether to respond yes or no to each trial. The subjects in the look and no-information conditions were then asked if they had noticed any relation between the star emblem and whether the plant bloomed. If they indicated that they believed the star emblem was related to flower blooming, they were asked to elaborate on how they thought it affected the plant and when during the experiment they had noticed this relation.

Results and Discussion

Two subjects, one in the look condition and one in the no-information condition, used the rating scales incorrectly and were excluded from the analyses. In each condition, the judgments after 80 trials from the remaining 20 subjects were analyzed.

Efficacy Ratings

Blue liquid. The subjects' causal ratings of the blue liquid, shown in Table 2, *decreased* linearly according to how much information the subjects had about the star emblem—from the no-information condition ($M = 22.4$) to the look condition ($M = 8.7$) to the tell condition ($M = -12.5$)—as is indicated by a linear trend analysis [$F(1,57) = 23.70, p < .0001$]. An omnibus F test showed an overall difference among the conditions [$F(2,57) = 12.03, p < .0001$]; a Fisher's PLSD revealed that the tell condition significantly differed from both the look and the no-information conditions. The tell condition therefore replicates the Simpson's paradox condition in the previous experiments, with ratings close to the conditional contingency. Importantly, the mean rating in the tell condition did not significantly differ from the conditional contingency [$t(19) < 1$], and it was significantly less than the unconditional contingency [$t(19) = 6.38, p < .0001$]. In the no-information condition, the rating is quite close to the blue liquid's *unconditional* contingency [$t(19) < 1$] (and to the rating in the equal condition of Experiment 2) and much greater than the blue liquid's conditional contingency [$t(19) = 5.86, p < .0001$]. Thus, because the subjects in this condition did not have a reason to believe that the star emblem was causal, they did not conditionalize the blue liquid's causal efficacy on it.

Star emblem. Although both the conditional and the unconditional contingencies for the star emblem were identical across the three conditions, the subjects' causal ratings *increased* linearly as the subjects had more information about the star emblem—from the no-information condition ($M = 37.4$) to the look condition ($M = 47.9$) to the tell condition ($M = 65.0$), as is indicated by a linear trend analysis [$F(1,57) = 7.97, p < .01$]. An omnibus F test revealed an overall difference among the conditions [$F(2,57) = 4.06, p < .05$]; a Fisher's PLSD showed

Table 3
Subjects' Report of Whether and When They Noticed a Relation Between the Star Emblem and Blooming

	Condition	
Report	Look (n = 20)	No-Information (n = 20)
Did they notice?		
Yes	19	12
No	1	8
When did they notice?		
Immediately	8	4
During the first ratings	5	0
After the first ratings	1	1
Never	1	8
Did not respond	5	7

that the tell condition differed significantly from the no-information condition.

Again, the tell condition replicates the Simpson's paradox condition in the previous experiments. The ratings in the other two conditions are less than those in the tell condition but greater than 0 [for the look condition, $t(19) = 7.76, p < .0001$; for the no-information condition, $t(19) = 4.69, p < .001$]. The ratings greater than 0 demonstrate that the subjects learned the contingency between the star emblem and the blooming despite not having a top-down theory of a causal mechanism. The subjects in the look condition knew to observe the star emblem and then could begin to recognize its effect. The subjects in the no-information condition would first need to notice the emblem and then notice that there was variation in the emblem's presence, before they could begin accruing the relevant covariation information.

Derived Contingencies

As in Experiment 2, the derived contingencies were consistent with the subjects' ratings, although the derived values were less extreme in each case. For the blue liquid, derived contingencies decreased linearly from the no-information condition ($M = 12.9$) to the look condition ($M = 4.3$) to the tell condition [$M = -5.5$; $F(1,57) = 10.17$, $p < .01$]. Derived contingencies significantly differed across conditions [$F(2,57) = 5.09$, $p < .01$]; a Fisher's PLSD revealed that the tell condition differed from the no-information condition. For the star emblem, the derived contingencies increased linearly from the no-information condition ($M = 15.9$) to the look condition ($M = 45.4$) to the tell condition [$M = 53.7$; $F(1,57) = 22.83$, $p < .0001$]. An omnibus F test indicated an overall difference among the conditions [$F(2,57) = 12.62$, $p < .0001$]; a Fisher's PLSD showed that both the tell and the look conditions significantly differed from the no-information condition. Again, these judgments were significantly correlated with the ratings [liquid, $r(62) = .64$, $p < .001$; emblem, $r(62) = .68$, $p < .001$].

Confidence Ratings

For the blue liquid, the subjects' confidence ratings in the tell ($M = 6.6$), look ($M = 6.2$), and no-information ($M = 5.7$) conditions did not significantly differ [$F(2,57) < 1$]. However, for the star emblem, the subjects in the tell condition were significantly more confident ($M = 7.5$) than those in the look ($M = 5.7$) and no-information ($M = 5.8$) conditions [$F(2,57) = 4.14$, $p < .05$, then a Fisher's PLSD]. Thus, the subjects were more confident in judging the star emblem when they had an explanatory causal mechanism (and, as a result, probably began accumulating information about the cause's efficacy from the beginning of the experiment).

Questionnaires

The questionnaires allowed us to learn more about how the subjects in the look and no-information conditions made their causal judgments (see Table 3). In response to the question of whether they had at any point noticed a relation between whether the pot was decorated and whether the flower bloomed, 19 out of 20 subjects in the look condition said that they had noticed, but only 12 out of the 20 subjects in the no-information condition said that they had noticed.

Examining those no-information subjects further, we see that for the star emblem, the mean rating for those who reported noticing a relation was higher ($M = 58.2$) than that for those who did not ($M = 6.2$); however, for the blue liquid, the mean rating for those who reported noticing a relation was lower ($M = 17.7$) than that for those who did not ($M = 29.4$). The interaction is significant [$F(1,18) = 25.5$, $p < .0001$].

To break down the analysis even further, for the subjects in the look and no-information conditions, we coded subjects as to when they noticed the relation. Table 3 shows their responses. We then ran a correlation between when they noticed the relation and their ratings for the star emblem and the blue liquid. For purposes of this correlational analysis, time of noticing was coded as follows: immediately = 1, during the first rating = 2, after the first rating = 3, never = 4. For the star emblem, we found that the earlier they caught on, the higher the rating [$r(28) = -.68$, $p < .0001$]. For the blue liquid, we found that the earlier they caught on, the lower the rating [$r(28) = .40$, $p < .05$]. Thus, as the subjects noticed the star emblem and accumulated information about it, their causal ratings for it increased. At the same time, ratings for the blue liquid decreased, suggesting that the subjects began to conditionalize on the now-recognized alternative cause.

Summary of Experiment 3

Experiment 3 illustrates both sides of Simpson's paradox—that identical information may be interpreted differently, depending on whether there is a reason to conditionalize on an existing alternative potential cause. In the tell condition, the subjects believed that the star emblem provided an alternative cause of blooming; therefore, the blue liquid's contingency was conditionalized on it. In the no-information condition, there is support from three different dependent variables (ratings, derived contingencies, and confidence) that at least some subjects

did learn something about the star emblem's causal efficacy in a strictly bottom-up manner. Because, on average, they did not believe the star emblem to be as causal as the subjects in the tell condition did or, perhaps, because they noticed the relation later and so did not have the same amount of information about the covariation between the two causes, the subjects in the no-information condition, on average, conditionalized less on the star emblem and judged the blue liquid to be more causal.

GENERAL DISCUSSION

In the case of two known potential causes of an effect, people use conditional rather than unconditional contingencies when evaluating the strength of those causes. They do so even when the alternative cause is odd but involves a plausible causal mechanism. This conclusion is supported by data from all three experiments. In both Experiments 1 and 2, causal ratings were higher for the blue liquid in the equal than in the Simpson's paradox conditions; in Experiment 3, when the subjects knew of the causal mechanism, we replicated the ratings for the Simpson's paradox conditions. These results suggest that the subjects were computing contingencies for the blue liquid while controlling for the potential alternative cause. Had the subjects been collecting merely marginal (unconditional) information, the Simpson's paradox and equal conditions (and the marginals-only condition in Experiment 1) would result in identical ratings. In addition to finding a significant difference in the ratings between the equal and the Simpson's paradox conditions, we also found that the subjects' ratings on a −100 to 100 scale were not significantly different from the *conditional* contingency (times 100) in those conditions. Derived contingencies showed the same overall pattern, with attenuated values.

The subjects showed conditionalization in both the summary table and the trial-by-trial formats. Mathematically, conditionalization requires people to be sensitive to and take into account the differing base rates in the cells. Such information is easily available in the summary table but needs to be extracted in the trial-by-trial format. Experiments 2 and 3 (tell condition) demonstrate that when subjects have a reason to believe that an event is causal, they can extract this information over repeated trials. The similarity in results between the two presentation formats (and the closeness of the ratings to the actual conditional contingency) is somewhat surprising in light of previous research demonstrating that causal judgments for single causes and effects tend to be better (i.e., closer to ΔP) when information is presented in a summary table rather than trial by trial (Kao & Wasserman, 1993; Shaklee & Mims, 1982; Ward & Jenkins, 1965). One might suppose that that difference would be exacerbated by the extra difficulty of keeping track of two causes in our tasks. A possible reason for the surprising current proficiency in the trial-by-trial task is that, rather than passively viewing the stimuli (as in those three articles cited above), subjects must make a prediction on each trial. The prediction technique has long been used in category-learning experiments (e.g., Medin & Schaffer, 1978) and is now often used in human causal learning experiments (e.g., Chapman & Robbins, 1990; Price & Yates, 1993). However, whatever the reason, these experiments demonstrate proficient conditionalization in a trial-by-trial task.[8]

We have also demonstrated that subjects seem to be both implicitly and explicitly aware that they need the frequency information about the covariation between the two causes to make causal judgments. In the summary table format of Experiment 1, those subjects who were given information only about the marginals were less confident of their causal judgments than were the subjects who got cell information; in addition, many subjects in the marginals-only condition explicitly asked for the cell information to be provided.

Therefore, in the case of two known or believed causes of an effect, subjects resolve apparent Simpson's paradoxes by (1) mathematically taking into account the differing base rates and (2) conceptually choosing to use the conditional rather than the unconditional contingency.

However, as Experiment 3 shows, judgments of causal strength are not always conditionalized on all covarying events. Rather, people only conditionalize when they have a reason to believe that a factor may be causally relevant to the effect; however, that reason may be a top-down causal theory or the bottom-up recognition of a covariation. In the tell condition in Experiment 3, which replicated the Simpson's paradox condition of Experiment 2, the subjects who were told about the star emblem mechanism conditionalized their ratings for the blue liquid on the star emblem. Thus, a top-down theory or a belief in a causal mechanism can drive conditionalization. However, we can also see a bottom-up effect. When some of the subjects in the look and no-information conditions in Experiment 3 began to notice the effect of the star emblem and collect information about it, not only did their ratings for the star emblem increase, but also their ratings for the blue liquid decreased (in accordance with conditionalization). Conditionalization was driven by noting the covariation between an alternative cause and the effect and then controlling for the alternative cause. Knowledge of a specific causal mechanism was not necessary; rather, the conditionalization strategy seems to have developed out of an interaction between bottom-up information and general theories of causality (Waldmann, 1996).

We do not know how strategic or automatic the detection, and use, of this covariation information is. In fact, we cannot be certain of exactly which information subjects are using to do these tasks. We have suggested that subjects are keeping track of three covariations: the covariation between each of the causes and the effect and the covariation between the two causes themselves. Cheng (personal communication, February 2000) has suggested

another possibility—namely, that once subjects have reason to believe that both factors are potentially causal, they assess each cause only in the absence of the other cause. Thus, for example, to evaluate the blue liquid, they need only consider the information in the two cells on the right of Figure 1, and to evaluate sunlight, they need only consider the information in the two cells at the bottom of Figure 1. Although this method may seem simpler than keeping track of three covariations or all four cells, it still involves (1) noting that a potential alternative cause exists, (2) judging that it is causal, and (3) electing to use some information (i.e., contingencies based on the absence of the alternative cause), rather than all of the information (i.e., the unconditional contingencies).

What would trigger the use of either of these conditionalizing strategies? In order to do these tasks, subjects may need to explicitly formulate and test hypotheses about how one cause behaves in the presence or absence of the other factor. Alternatively, perhaps all subjects need is a reason to begin keeping (automatic) track of the conditional frequencies. We suspect that all is not automatic. For example, under some conditions of divided attention, even in the presence of a top-down theory, conditionalizing does not occur (Goedert & Spellman, 2000). Such questions are being addressed in further experiments. What we do know is that, under certain conditions, people do not see Simpson's paradox as a paradox and that, under some of those conditions, they choose to resolve it in a mathematically sophisticated way that involves using conditional contingencies.

REFERENCES

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, **54**, 299-352.

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

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209-254). San Diego: Academic Press.

Cartwright, N. (1979). *How the laws of physics lie*. Oxford: Oxford University Press, Clarendon Press.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 837-854.

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

Cheng, P. W. (1993). Separating causal laws from casual facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The psychology of learning & motivation: Advances in research and theory* (Vol. 30, pp. 215-264). San Diego: Academic Press.

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

Flexser, A. J. (1981). Homogenizing the 2 × 2 contingency table: A method for removing dependencies due to subject and item differences. *Psychological Review*, **88**, 327-339.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 513-525.

Goedert, K. M., & Spellman, B. A. (2000). *Controlling for alternative causes may require attention; discounting does not*. Unpublished manuscript, University of Virginia.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, **39**, 1372-1388.

Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, **87**, 398-410.

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

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral & Brain Sciences*, **19**, 1-53.

Martin, E. (1981). Simpson's paradox resolved: A reply to Hintzman. *Psychological Review*, **88**, 372-374.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

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, & Cognition*, **19**, 1398-1410.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Vintage Books.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Price, P. C., & Yates, J. F. (1993). Judgmental overshadowing: Further evidence of cue interaction in contingency judgment. *Memory & Cognition*, **21**, 561-572.

Price, P. C., & Yates, J. F. (1995). Associative and rule-based accounts of cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1639-1655.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Schaller, M. (1992a). In-group favoritism and statistical reasoning in social inference: Implications for formation and maintenance of group stereotypes. *Journal of Personality & Social Psychology*, **63**, 61-74.

Schaller, M. (1992b). Sample size, aggregation, and statistical reasoning in social inference. *Journal of Experimental Social Psychology*, **28**, 65-85.

Schaller, M., & O'Brien, M. (1992). "Intuitive analysis of covariance" and group stereotype formation. *Personality & Social Psychology Bulletin*, **18**, 776-785.

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

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

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology*, **48A**, 257-279.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B*, **13**, 238-241.

Spellman, B. A. (1993). *The construction of causal explanations*. Unpublished doctoral dissertation, University of California, Los Angeles.

Spellman, B. A. (1996a). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science*, **7**, 337-342.

Spellman, B. A. (1996b). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 167-206). San Diego: Academic Press.

Spellman, B. A. (1996c). The implicit use of base rates in experiential and ecologically valid tasks. *Behavioral & Brain Sciences*, **19**, 38.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131. [Reprinted in D. Kahneman, P. Slovic, & A. Tversky (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases* (pp. 3-20). New York: Cambridge University Press.]

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under un-*

certainty: Heuristics and biases (pp. 153-162). New York: Cambridge University Press.

WAINER, H. (1986). Minority contributions to the SAT score turnaround: An example of Simpson's paradox. *Journal of Educational Statistics*, **11**, 239-244.

WALDMANN, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol. 34. Causal learning* (pp. 47-88). San Diego: Academic Press.

WALDMANN, M. R., & HAGMAYER, Y. (1995). When a cause simultaneously produces and prevents an effect. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 425-430). 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.

WASSERMAN, E. A., ELEK, S. M., CHATLOSH, D. L., & BAKER, A. G. (1993). Rating causal relations: Role of probability in response outcome contingency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 174-188.

WHITE, P. A. (1995). Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory & Cognition*, **23**, 243-254.

NOTES

1. The two conditional contingencies do not have to be equal. See Spellman (1996b) for a discussion of the mathematical properties of conditional contingencies.

2. Unconditional contingencies can remain constant while the marginals are varied; however, we kept the marginals constant in order to avoid outcome density effects (Wasserman, Elek, Chatlosh, & Baker, 1993; see Cheng, 1997, pp. 390 392, for an explanation).

3. Note that it does not make sense to compare the marginal-only result to the Simpson's paradox condition. If subjects were not assuming equal frequencies in the cells, there would be many ways to assign the frequencies; the assignment need not be that used in our particular Simpson's paradox condition.

4. If anything, the ratings for the star emblem should be highest in the Simpson's paradox condition, where its conditional contingency is .67, as opposed to .60 in the equal condition and indeterminable in the marginals-only condition.

5. Note that the former ratings are judgments of the blue liquid's sufficiency and the latter are judgments of its necessity. Thanks to Denise Dellarosa Cummins for pointing this out.

6. Waldmann and Hagmayer (1995, Experiment 2) do not have a baseline of comparison (as in our equal condition). Those subjects who are not given a causal theory may still be conditionalizing on the covarying alternative potential cause.

7. But see the discussion on pp. 246 248 about how prior beliefs may act as covariation information.

8. In Waldmann and Hagmayer (1995), the list containing all of the information remained in front of the subjects as they made their judgments. Some of Schaller's experiments were done in a trial-by-trial manner, and conditionalization did occur, but the extent of it could not be evaluated because there was no analog to the equal (baseline) condition.

APPENDIX
Simpson's Paradox in the Baseball Example

This study used the baseball example described in the introduction to discover how much information subjects need before they will recognize the possibility of a Simpson's paradox in a well-known domain. Subjects were given information about the batting averages and frequencies of two players in each half of the baseball season.

METHOD

Subjects and Procedure

The subjects were 225 University of Texas undergraduates who participated in partial fulfillment of a course requirement. This experiment was administered in a booklet along with other short reasoning tasks.

Materials and Design

The subjects rated their baseball knowledge on a scale from 0 (*I don't know anything about baseball*) through 4 (*I am an avid fan*). They then read about two hypothetical baseball players' performances in the two halves of the baseball season. All the conditions conveyed the information that Player A's average was higher in both halves of the season; we manipulated *how* that information was conveyed (see Table 1A). First, the information about the batting averages was in the form either of relative averages (e.g., A has a higher average than B) or of actual averages (e.g., A's average is .400, B's average is .300). Second, crossed with that variable, the subjects either did or did not get information about the frequency of at bats for each player in each half of the season. That created a 2×2 design. In addition, for the combination in which both actual averages and frequencies were given, the information was presented in two different ways, either stating the average and then the frequency of at bats (e.g., A's batting average is .400; A was at bat only 10 times) or stating the average and then the frequency of both hits and at bats (e.g., A's batting average is .400 [4 for 10]).

After reading the information, all the subjects were asked, "Which player has a higher full season average?" They were instructed to circle one of the following: (a) Player A, (b) Player B, or (c) it depends. They were then asked to briefly explain their answers.

RESULTS AND DISCUSSION

Answers

In each of the first three conditions, an overwhelming majority of subjects picked Player A despite the fact that the correct answer was *it depends* (see Table A1). In particular, in the first two conditions, in which the subjects only got either relative or actual averages, hardly any subjects gave any answer other than Player A. The pattern of answers given in these two conditions did not differ [$\chi^2(2, N = 88) = 1.93$, n.s.]. In the third condition, which added frequency information to the relative average, the subjects were more likely to answer *it depends* than the subjects who only had the relative average information [$\chi^2(2, N = 87) = 7.09, p < .05$]. Thus, the frequency information seemed to be cuing them into the idea that frequencies could be important.

In the last two conditions, a substantial number of subjects picked Player A despite the fact that there was enough information for them to determine that the correct answer was Player B. However, correct performance was better in the last two conditions than in the first three conditions [48/91 vs. 14/134 correct; $\chi^2(2, N = 225) = 48.58, p < .001$].

Note two differences between the later and the earlier conditions: First, the later ones did, in fact, give all the necessary data (and as a result, the correct answers were different). But second, by giving all the data, these conditions conveyed the idea that both actual average and frequency were important. Subjects do not usually seem to spontaneously think that frequency is important, but they are more likely to acknowledge its importance

Table A1
Information Provided and Percentage of Subjects in Each Condition Giving Each Answer in the Baseball Experiment

Condition	*n*	Information				Subjects' Answers (%)		
		First half of the season		Second half of the season				
		Player A	Player B	Player A	Player B	Player A	Player B	Depends
1. Relative average (as in text example)	41	A has a higher average than B		A has a higher average than B		93	2	**5**
2. Actual average	47	A's average is .400	B's average is .300	A's average is .250	B's average is .200	89	9	**2**
3. Relative average plus frequency	46	A has a higher average than B, but A was at bat only 10 times, whereas B was at bat 100 times	A has a higher average than B, but A was at bat 100 times, whereas B was at bat only 10 times			76	0	**24**
4. Actual average plus frequency (at bats only)	45	A's average is .400; A was at bat only 10 times	B's average is .300; B was at bat 100 times	A's average is .250; A was at bat 100 times	B's average is .200; B was at bat only 10 times	40	**56**	4
5. Actual average plus frequency (hits out of at bats)	46	A's average is .400 (4 for 10)	B's average is .300 (30 for 100)	A's average is .250 (2 for 10)		65	**28**	7

Note—Bold numbers indicate the correct answer given the information provided.

once it is mentioned (as the difference between Conditions 1 and 3 suggests).

Explanations and Knowledge

When the subjects were asked to explain their answers, 50% either wrote something like "A is higher in both halves of the season" or mathematically averaged the batting averages, showing that A was higher. Each of these answers ignores the importance of frequency-per-half information.

Even when we divided subjects into high-knowledge (rating > 2.5) and low-knowledge (rating < 2.5) groups, we found that high-knowledge subjects were not more likely to give the correct answer of *it depends* in Conditions 1 3 (i.e., they did not see the importance of the frequencies of at bats in each half of the season). However, in Conditions 4 and (marginally) 5, in which the subjects could figure out the correct answers, high-knowledge subjects were proportionately more likely to use the correct math to get the solution than were low-knowledge subjects.

CONCLUSION

In the familiar domain of baseball statistics, the subjects did not easily recognize the possibility of a Simpson's paradox. Granted, there are several things working against subjects' recognizing that there might be a problem. It seems that in baseball, if Player A's average is higher in both halves of the season, then usually Player A's overall average *would* be higher. The mathematical precondition of widely differing frequencies is probably not often met. In addition, there is really no conceptual cue that season half should matter; there is no obvious causal link between season half and batting average. Thus, under these two conditions, when there is not enough evidence to make the full computation, even experts fail to recognize the paradox that differing frequencies could create.

(Manuscript received August 9, 1999;
revision accepted for publication August 16, 2000.)