Memory & Cognition
1995, 23 (4), 510-524

Causal inferences as perceptual judgments

JOHN R. ANDERSON and CHING-FAN SHEU
Carnegie Mellon University, Pittsburgh, Pennsylvania

We analyze how subjects make causal judgments based on contingency information in two paradigms. In the discrete paradigm, subjects are given specific information about the frequency a, with which a purported cause occurs with the effect; the frequency b, with which it occurs without the effect; the frequency c, with which the effect occurs when the cause is absent; and the frequency d, with which both cause and effect are absent. Subjects respond to $P_1 = a/(a+b)$ and $P_2 = c/(c+d)$. Some subjects' ratings are just a function of P_1, while others are a function of $\Delta P = P_1 - P_2$. Subjects' post-experiment reports are accurate reflections of which model they use. Combining these two types of subjects results in data well fit by the weighted ΔP model (Allan, 1993). In the continuous paradigm, subjects control the purported causes (by clicking a mouse) and observe whether an effect occurs. Because causes and effects occur continuously in time, it is not possible to explicitly pair causes and effects. Rather, subjects report that they are responding to the rate at which the effects occur when they click versus when they do not click. Their ratings are a function of rates and not probabilities. In general, we argue that subjects' causal ratings are judgments of the magnitude of perceptually salient variables in the experiment.

There has been a long history of discussion about human causal inference. Philosophers (e.g., Cohen, 1981; Henle, 1962; Hume, 1740/1938; Mill, 1843/1974; Popper, 1972; Suppes, 1970) have been concerned with when we can be justified in inferring a causal relationship, and psychologists (e.g., Jenkins & Ward, 1965; Kahneman & Tversky, 1972; Lipe, 1982; Michotte, 1963; Nisbett & Ross, 1980; Peterson & Beach, 1967; Shultz, 1982; Smedslund, 1963; Tolman & Brunswik, 1935) have been concerned with when people actually infer a causal relationship. The goal of this paper is to present some data on causal inferences based on contingency information, which is one domain where the rationality of causal inference has been addressed. We will look at judgments first in what we call the discrete paradigm and later in what we call the continuous paradigm. A typical discrete problem might be to determine the relationship between jogging and backache given data that can be understood in terms of a 2×2 contingency table such as Table 1. The subject is told of cases where the patient jogs and has a backache (a), jogs and has no backache (b), does not jog and has a backache (c), and neither jogs nor has a backache (d).

In Table 1, the events are classified according to whether the purported cause is present or absent and the effect occurs or does not occur. The values a through d are the frequencies in these cells. The subject's task is to go from these frequencies to an inference about whether there is a causal relationship or not. In most experiments, including the one we will report, subjects do not see such a table but, rather, experience the various event combinations with the specified frequencies. One question that has been asked is for what patterns of these four numbers should subjects infer a causal relationship and for what patterns of numbers do they, in fact, infer a relationship.

One can calculate the $P_1 = a/(a+b)$, which is the proportion of times the effect occurs in the presence of the cause, and $P_2 = c/(c+d)$, which is the proportion of times the effect occurs in the absence of the cause. Allan (1980, 1993; Allan & Jenkins, 1983) has suggested that the appropriate measure of dependency is

$$\Delta P = P_1 - P_2.$$

This same statistic has been used by Wasserman (1990), and a variant of it, called focal-set ΔP, has been advocated as a model of human causal inference by Cheng and Novick (1992).

One problem with the ΔP rule is that subjects are not equally sensitive to all cells, as this rule would imply. Therefore, more successful fits to the data are reported using the weighted ΔP rule:

$$w_0 + w_1 P_1 - w_2 P_1.$$

These fits are often very good in absolute terms, and it is typically found that $w_2 < w_1$.

The goodness of fit of a weighted ΔP rule is subject to multiple interpretations. The straightforward one is that subjects are calculating the probabilities of the effect in the presence and absence of the cause and then weighting them to come up with their causal judgment. This has struck some researchers as implausible, and

This research was supported by Contract N00014-91-J-1597 from the Office of Naval Research. We would like to thank Jon Fincham for programming the four experiments and Marsha Lovett and Lael Schooler for their comments on the paper. Correspondence should be directed to J. Anderson, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 (e-mail: ja+@cmu.edusb.).

Table 1
A Typical 2 × 2 Contingency Table for Causal Estimation

	Effect Present (e.g., back pain)	Effect Absent (e.g., no back pain)
Cause present (e.g., jogging)	a	b
Cause absent (e.g., no jogging)	c	d

subjects often fail to report engaging in such conscious calculations. It could be that the weighted ΔP reflects some implicit strength that builds up between the cause and the effect. Chapman and Robbins (1990) and Wasserman, Elek, Chatlosh, and Baker (1993) have shown that, under certain assumptions, the Rescorla–Wagner (Rescorla & Wagner, 1972) rule leads to associative strength proportional to a weighted ΔP. Shanks (1987) has shown that the growth of causal ratings with exposure corresponds to the predictions of the Rescorla–Wagner theory. Allan (1993) has reviewed a number of correspondences between effects in the conditioning literature and the ΔP rule.

The ΔP rule, under any of these interpretations, has problems as a normative model of human causal inference. One problem is that it shows no sensitivity to sample size, and any normative model would hold that the strength of belief in a causal relationship should increase as sample size increases. In the limit, as sample size increases, a normative model should assign a certainity to the existence of a causal relationship for any positive ΔP. Allan (1980) has suggested that, normatively, the significance of ΔP should be assessed by a chi-square test, but there is no evidence subjects are doing this. The Rescorla–Wagner interpretation, because of the learning process, shows some sensitivity to sample size because it takes a number of trials for conditioning to asymptote to ΔP. However, beyond enough data to estimate ΔP accurately, the Rescorla–Wagner model is not sensitive to sample size.

Besides its insensitivity to sample size, the weighted ΔP rule might also seem nonnormative because of its unequal weighting of the cells. The chi-square logic would seem to imply that subjects should be equally sensitive to all cells. Many researchers have noted that subjects are most sensitive to variations in a and least sensitive to variations in d. In some experiments, subjects appear totally insensitive to d, while in others, subjects are just less sensitive to d (Arkes & Harkness, 1983; Crocker, 1981; Shaklee & Tucker, 1980; Wasserman, Dorner, & Kao, 1990). This insensitivity to the cell reflecting joint absence has led to the characterization of humans as nonrational in their causal inference.

Schustack and Sternberg (1981) proposed a model that was sensitive to sample size and that could reflect the unequal cell weighting. They had a great deal of success fitting their data sets using a simple linear model of the form

$$w_0 + w_1 a + w_2 b + w_3 c + w_4 d.$$

This linear model predicts that causal judgment is a weighted function of the individual cell frequencies, and so makes judgment sensitive to sample size and also can accommodate differential sensitivity to frequencies in different cells. Although the differential sensitivity to cell frequency has been well documented in the discrete paradigm, the issue of the overall influence of sample size has not been carefully studied in the discrete paradigm.[1] So, it is not clear whether the ΔP or the linear weighting model would provide a better characterization of the data.

The Simple Bayesian Model

Anderson (1990) proposed a Bayesian analysis for what normative causal inferences should be. This analysis proposed that subjects compared the data in a 2×2 contingency table with a model of what a causal relationship was like. One model is that there should be a certain probability, p_C, of an effect in the presence of a cause and another probability, p_A, of the effect in the absence of the cause. The overall likelihood of the data is then the product of the probabilities of two independent sequences of Bernoulli trials (i.e., the sequences summarized by the first and second rows of Table 1). The first one is the probability of a successes and b failures, when the cause is present, given that the probability of success is p_C. The second one is the probability of c successes and d failures, when the cause is absent, given that the probability of success is p_A. That is,

$$P(D|H) = p_C^a(1 - p_C)^b\, p_A^c(1 - p_A)^d. \tag{1}$$

This is the likelihood of the data (D) under the hypothesis (H) of a causal relationship. This needs to be compared with the likelihood of the data under the hypothesis that no cause was identified ($\bar{H}$). In this case, there should be a base probability, p_N, governing the occurrence of the effect whether the purported cause is present or not, and the two rows of Table 1 can be collapsed to give $a + c$ successes and $b + d$ failures. Thus, the likelihood of the data under no identifiable cause is

$$P(D|\bar{H}) = p_N^{a+c}(1 - p_N)^{b+d}. \tag{2}$$

Our inference about whether there is a causal relationship should be a function of the odds of a causal relationship given the data:

$$\text{Odds}(H|D) = \frac{P(H|D)}{P(\bar{H}|D)}$$

$$= \frac{P(H)}{P(\bar{H})} \cdot \frac{p_C^a(1 - p_C)^b p_A^c(1 - p_A)^d}{p_N^{a+c}(1 - p_N)^{b+d}}, \tag{3}$$

where $P(H)$ is the prior probability of a causal relationship and $P(\bar{H}) = 1 - P(H)$. Anderson (1990) presents a more complex version of this model which involves Bay-

esian assumptions about prior distributions of probabilities. This model, rather than assuming fixed priors for p_C, p_A, and p_N, assumes prior distributions of such probabilities. A somewhat similar idea is described by Fales and Wasserman (1992). However, we have found little difference between the models with fixed priors and distributions of priors. Therefore, we will consider only the simpler model with fixed priors.

Schustack and Sternberg Experiment

The data of Schustack and Sternberg (1981) provide an example for applying these models. We will apply it to the abstract condition of their third experiment.[2] They presented their subjects with 60 test items of the form

S	W	N	T	→	R
−Q	N	Z	S	→	−R
T	W	S	Z	→	R
T	−S	−Q	−N	→	−R
			S	→	R

The first four lines are abstract characterizations of possible causal events. The first says when S, W, N, and T were present R occurred, and the second says when Q was absent and N, Z, and S were present R did not happen, and so forth. The last item is the one subjects are asked to judge—in this case, to judge the likelihood that S leads to the outcome of R on a rating scale from 0 to 100. (Subjects assigned an average value of 34.7 to the example above.) Such problems can be described in terms of the values of a, b, c, and d. In this example, $a = 2$, $b = 1$, $c = 0$, and $d = 1$, where these reflect the co-occurrence pattern of S and R. The 60 problems used by Schustack and Sternberg involved substantial variations in the values of a through d and so allow us to put the model to a serious test.

Schustack and Sternberg (1981) fit their linear model to these data, which used six variables—the values of a through d, a single measure of the strength of alternative causes, and a regression constant. This model fit the data quite well—with an R^2 of .90. Corresponding to other findings (e.g., Smedslund, 1963), the regression coefficients are largest for a and least for d. From this, Schustack and Sternberg concluded that "subjects are not optimally rational" (p. 119). In addition, they reported that attempts to implement a Bayesian variable, along with the variables of their proposed model, did not produce encouraging results. Therefore, "subjects were not using a Bayesian approach in their evaluation" (p. 113). While Schustack and Sternberg's linear model appears to provide a good quantitative account of subjects' performance in the inference task, their conclusion about the subjects' rationality is not warranted. We will show that the simple Bayesian model can fit their data quite well.[3]

Fit of the Simple Bayesian Model

The subjects in the Schustack and Sternberg paradigm gave responses on a 0–100 scale, which suggests that they treated them as probabilities. Therefore, we used the following response rule:

$$\text{Response} = 100 \times \frac{\text{odds}(H|D)}{1 + \text{odds}(H|D)}, \tag{4}$$

where odds$(H|D)$ is calculated according to Equation 3. The resulting R^2 is .90. The four parameter values were $P(H) = .29$, $p_C = .63$, $p_N = .41$, and $p_A = .27$. The correlations between the predictions of this model and the values of a, b, c, and d were .87, −.61, −.57, and .17, respectively. Thus, just as subjects do, the rational model can prescribe that the four variables should receive differential weighting. It is easier to understand why the simple Bayesian model produces the differential correlation with a versus d if one examines the log of the odds formula in Equation 3:

$$\begin{aligned}\log\left(\frac{P(H|D)}{P(\bar{H}|D)}\right) = {} & \log\left[P(H) - P(\bar{H})\right] \\ & + a\left[\log(p_C) - \log(p_N)\right] \\ & + b\left[\log(1 - p_C) - \log(1 - p_N)\right] \\ & + c\left[\log(p_A) - \log(p_N)\right] \\ & + d\left[\log(1 - p_A) - \log(1 - p_N)\right]. \end{aligned} \tag{5}$$

Thus, there is a linear relationship between log odds and the variables a through d. In fact, Equation 6 asserts that the simple Bayesian model is equivalent to a linear regression model on log odds. The four coefficients, $\log(p_C) - \log(p_N)$, $\log(1-p_C) - \log(1-p_N)$, $\log(p_A) - \log(p_N)$, and $\log(1-p_A) - \log(1-p_N)$, assigned to variables a through d depend only on the three parameters, p_C, p_A, and p_N, of the simple Bayesian model. To the extent that $\log(p_C) - \log(p_N)$ is large and $\log(1-p_A) - \log(1-p_N)$ is small, subjects should weight a more than d. Intuitively, p_A, which is the probability of an effect when the cause is absent, should not be very different from p_N, which is the base probability. On the other hand, there should be a large difference between p_C, which is probability when the effect is present, and p_N. These intuitions imply that every joint occurrence of cause and effect will provide more evidence for the cause than every joint absence.

Thus, the prescription of rationality does not require equal weighting of a and d. The subject is comparing the causal model against the model of no known cause. It is reasonable to assume that p_C is high and p_N and p_A are low and relatively equal. Thus, the coefficient for a is expected to be large and for d small. To the extent that the base probability of an effect is the same as the probability with cause absent ($p_N = p_A$), the variable d does not provide discriminating data.

A symmetric treatment of a and d is required if we take the chi-square view of trying to determine whether

there is a significant difference in frequencies in the presence of a purported cause versus its absence. However, this is not what subjects were asked to do in Schustack and Sternberg's experiment. Rather, they were asked whether the described information better supported a causal relationship. In that framework, variations in the variable d should be treated as less important than variations in a. This is just what subjects did. Their behavior can be predicted extremely well by a meaningful set of parameters underlying a rational analysis of how they should perform the task.

As a final comment, Equation 5 shows that rather sophisticated Bayesian inference can be achieved by a very simple response rule. One might regard Equation 5 as describing how strength of association (interpreted as log odds) should change with the frequencies a, b, c, and d. Basically, each such event would increase or decrease the strength by an amount that depended on the parameters p_C, p_A, and p_N. While such a "rational" model is computationally plausible, it does not necessarily describe what subjects do. This is the issue that we explored in the following series of experiments.

EXPERIMENT 1

There is little empirical basis upon which to choose between the Schustack and Sternberg linear model and the simple Bayesian model in terms of their ability to fit the data from the Schustack and Sternberg experiment. Both predict these data with R^2s of .90 although the simple Bayesian model has two fewer parameters. However, they differ fundamentally in the relationship they predict between the independent variables, a to d, and the dependent judgment measures. The Schustack and Sternberg model predicts that the actual judgments should bear a linear relationship to a through d while the simple Bayesian model predicts that the log-odds transforms of the judgments should bear a linear relationship to a through d, and hence the actual judgments should be a negatively accelerated function of these variables. Unfortunately, the values of a through d in Schustack and Sternberg were not manipulated over a sufficiently wide range to allow these two possibilities to be distinguished. Also because of this restricted range, the experiment did not allow a test of how sensitive subjects are to sample size. Both the Bayesian model and the linear model differ from the ΔP model on this score. The motivation of this experiment was to present subjects with problems that allowed the Bayesian model to be discriminated from both the linear model and the ΔP model. The ΔP model could not be tested against the Schustack and Sternberg data because P_1 and P_2 were not always defined. We wanted to create a data set that would also allow us to test the ΔP model.

Method

In this experiment, subjects were asked to evaluate the likelihood of a drug's causing side effects in the treatment of a fictitious disease. Four types of information corresponding to the different cells in a 2×2 contingency table were available: the joint presence of the side effects and drug treatment, absence of the side effects

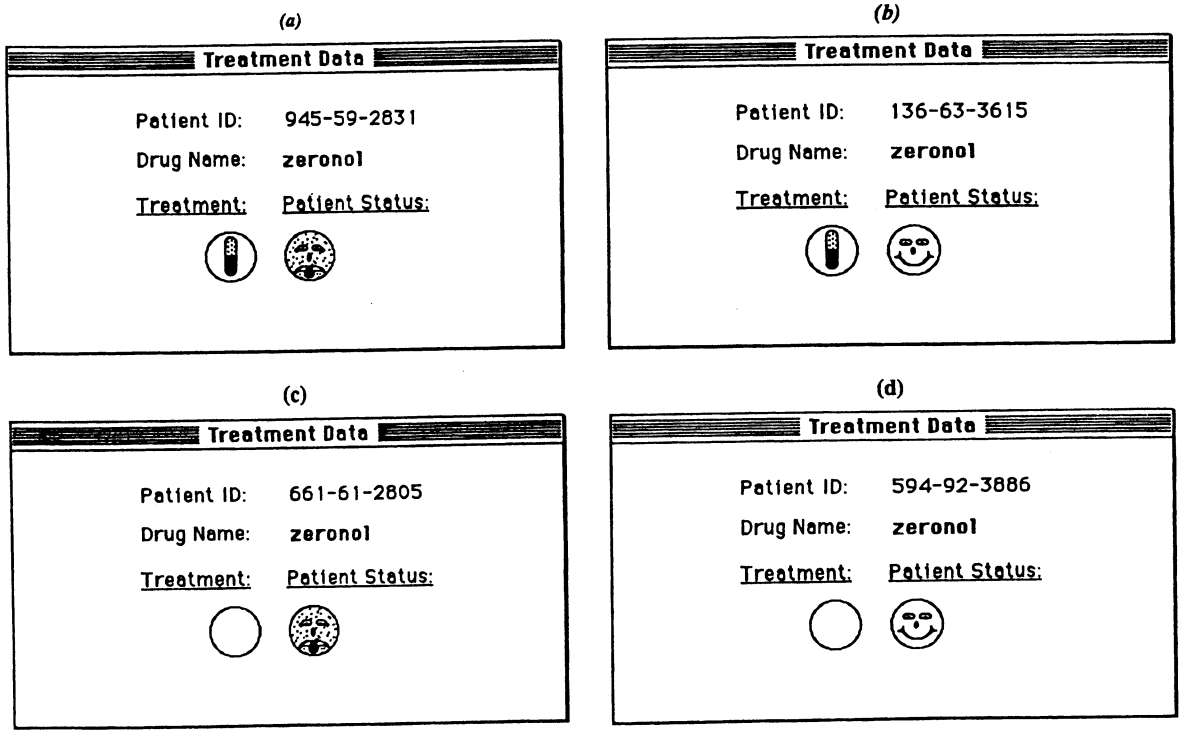

Figure 1. Examples of stimuli used in Experiment 1. (a) Cause present and effect present; (b) cause present and effect absent; (c) cause absent and effect present; (d) cause absent and effect absent.

given the drug treatment, presence of the side effects given no drug treatment, and the joint absence of drug treatment and side effects. The subjects were asked to judge a number of problems, and each problem involved a sequence of instances of these four information types. The frequencies of each information type varied from problem to problem. At the end of a problem, the subjects were asked to enter a number from 0 to 100 that best reflected their judgment of the drug's causing the side effects.

Subjects. Forty graduate and undergraduate students were recruited through an ad on the Carnegie Mellon University computer market bulletin board. They were each paid $7 for participating in the 1-h experiment. One subject opted for and received, instead, a credit for fulfilling a course requirement. There were 27 male and 13 female subjects, of whom 23 had taken statistics courses. Neither sex nor the prior statistics course interacted with the results, so we collapsed the results for all subjects.

Apparatus. A Mac IIci computer was used to control the stimulus display and collect data from the subjects. The display was a 5×3 in. window at the center of an Apple 21-in. monochrome monitor.

Procedure. Subjects were presented with 2 warm-up problems and then 80 experimental problems, which they were to judge for causal efficacy. The information for each problem was presented in pictorial format, as shown in Figure 1. A problem consisted of the presentation of a drug name and a flashing fixation point, a sequence of patient outcomes, like those in Figure 1, and then a display to solicit response entry. A different drug name was chosen for each problem and a different patient identification number was selected in each information display. The order of problems and the order of the information displayed within a problem were randomized for each subject. Each display for each patient stayed on the screen for 1 sec, and there was a 1/4-sec interval between patient displays.

At the beginning of the experiment, the subjects were given the following information to read:

> A new family of drugs to treat Lafuma's syndrome has been approved for clinical trials. The drugs were assigned code names and distributed to 80 hospitals nation-wide.
>
> Two common side effects of the drugs are skin rash and the growth of tongue moss. Unfortunately, these are also two of the symptoms that patients with Lafuma's syndrome have. The pharmaceutical company is interested in knowing what doctors think about the chance of each of the different drugs causing the side effects. They collect results from each of the hospitals and classify the outcome of each patient in one of the four possible ways:
>
> 1. the drug was administered and the side effects were present.
> 2. the drug was administered and no side effects were observed.
> 3. the drug was not administered but the side effects were present.
> 4. the drug was not administered and the side effects were not observed.
>
> The doctor at the pharmaceutical company examined the outcome case by case. After completing the results from one hospital, the doctor gave a rating from 0 to 100 to indicate how likely the drug used in that hospital is the cause of side effects in patients. Then data from another hospital is reviewed. Now, suppose you are the doctor working at the company. What will your judgments be?

The 0–100 scale was used to replicate Schustack and Sternberg.

Design. The 80 problems are described in Table 2. The problems are divided into four sets. Each set was intended to evaluate the effect of one of the four information types (*a* to *d*). Along each row of a set, the values of three of the information types are held constant while the value of the fourth varies over the values of 1, 3, 5, 9, and 15.

Table 2
Mean Rating of Drug Data over all Subjects

		Value of Variable				
	a, *b*, *c*, *d*	1	3	5	9	15
vary *a*	_,1,1,1	39	62	68	66	77
	_,3,3,3	29	42	50	61	70
	_,5,5,5	19	38	40	51	66
	_,1,3,5	38	61	62	77	81
vary *b*	1,_,1,1	41	31	18	14	9
	3,_,3,3	54	39	35	23	19
	5,_,5,5	60	48	46	30	26
	1,_,3,5	43	25	20	13	9
vary *c*	1,1,_,1	43	36	29	26	20
	3,3,_,3	47	41	37	35	28
	5,5,_,5	50	44	43	35	32
	5,3,_,1	53	48	39	33	37
vary *d*	1,1,1,_	47	40	52	50	49
	3,3,3,_	43	37	43	49	49
	5,5,5,_	35	41	42	46	50
	5,3,1,_	62	60	57	59	62

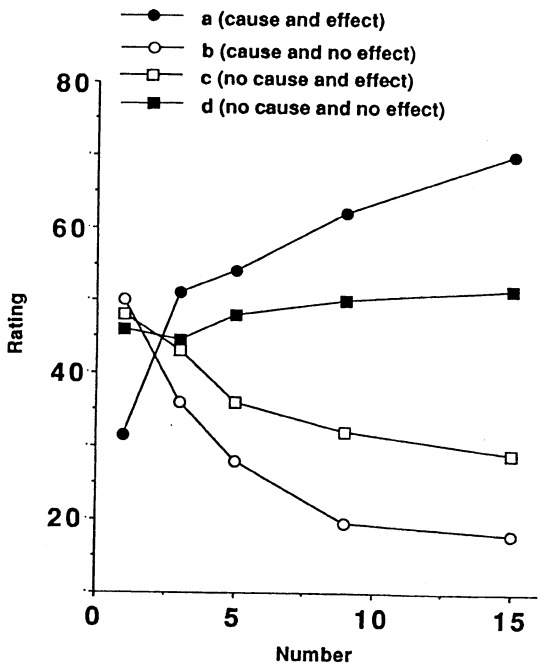

Figure 2. The effect of variation in the variables *a* through *d* on judged efficiency of the cause.

Results

Table 2 presents the mean ratings of subjects for each of the 80 conditions. Figure 2 shows how judgments varied as a function of *a* through *d*, each with the other three variables held constant. Separate analyses of variances were performed on the submatrices in Table 2 for the variables *a*–*d* using as factors the rows and columns. There were significant effects of each column variable [$F(4,156) = 119.8, p < .001$, for *a*, $MS_e = 337.2$; $F(4,156) = 98.5, p < .001$, for *b*, $MS_e = 292.4$; $F(4,156) = 33.7, p < .001$, for *c*, $MS_e = 282.4$; $F(4,156) = 5.1, p < .001$, for *d*, $MS_e = 325.3$].[4] As in other research, the largest effect is for *a* and smallest is for *d*. The functions do appear to be

negatively accelerated as would be predicted by the simple Bayesian model. As a test of whether the functions were indeed negatively accelerated, we tested whether or not the change from 1 to 5 was greater than the change from 9 to 15. The Bayesian model predicts a greater change, while the linear regression model predicts a smaller change. The increase from 1 to 5 was larger for each variable and significantly larger for all variables except d[$t(156) = 4.95$, $p < .001$, for a; $t(156) = 5.63$, $p < .001$, for b; $t(156) = 3.11$, $p < .01$, for c; $t(156) = .17$, for d].

Sample size neglect. While the evidence for a curvilinear effect of a through d supports the Bayesian model over the Schustack and Sternberg model, there remains the question of whether either is correct in its prediction that subjects' judgments should become more extreme with sample size. The data include some conditions which allow fairly direct assessments of this issue. For instance, there are three cases in which the value of a is three times as large as the values of b, c, and d—3, 1, 1, 1; 9, 3, 3, 3; and 15, 5, 5, 5. Similar comparisons are possible for b, c, and d. Figure 3 plots these comparisons as a function of sample size. As can be seen, there was only a weak trend for subjects to become more extreme in their opinions even though sample size was increasing by a factor of 5. The effect for a is not significant [$F(2,78) = 1.61$], that for b is significant [$F(2,78) = 4.02$, $p < .05$], that for c is not [$F(2,78) = 1.03$], and that for d is [$F(2,78) = 4.41$, $p < .05$]. Thus, it seems that the data would be more consistent with something like the ΔP model, which is insensitive to sample size. According to Equation 5, the difference in log-odds between the most extreme conditions in Figure 3 should increase by a factor of 5. If we treat subjects' judgments as probabilities and convert these to log-odds, the difference in the judgments for the a and b cases only increases from 1.20 to 1.67. Subject neglect of sample size in other decision-making paradigms has also been noted by a number of other researchers (e.g., Tversky & Kahneman, 1974).

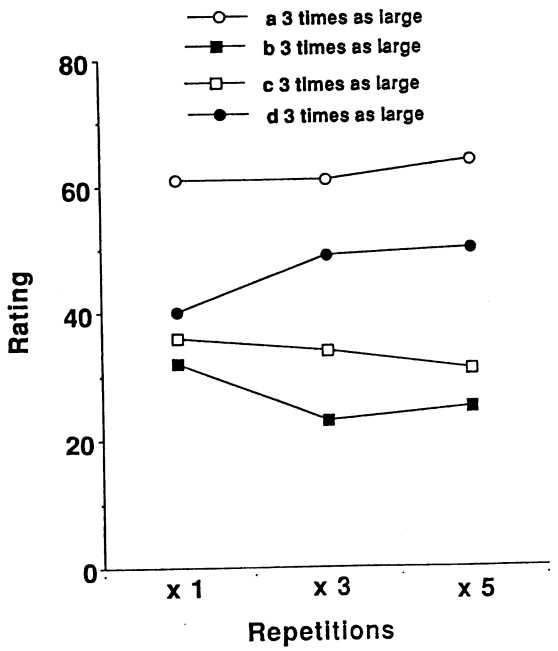

Figure 3. Effect of number of repetitions of data on judgment—predicted indicated by closed symbols and observed indicated by open symbols.

Shanks (1985) reported an experiment in the discrete paradigm in which subjects did show sensitivity to sample size. However, his data showed relatively little effect beyond 5–10 observations, while only 20% of our cases involved 10 or fewer observations. Also, he queried subjects after every five trials as to what their estimates were, whereas we asked only for the final estimate. Thus, it is not clear that the weak effects we find are any different from his. The larger effects that he found may be due to the requirement for repeated evaluations which might encourage subjects to become more extreme in their reports.

Model fits. We fit four models to the data which confirm the impressions obtained in the preceding analyses. The linear Schustack and Sternberg model (with parameters of $w_0 = 40.8$, $w_1 = 3.4$, $w_2 = -2.5$, $w_3 = -1.5$, and $w_4 = .08$) fit with an R^2 of .86. The Bayesian model (with parameters of prior = .40, $p_C = .50$, $p_N = .42$, $p_A = .40$) fit with an R^2 of .89. The weighted ΔP model (with parameters of $w_0 = 19$, $w_1 = 72$, and $w2 = 27$) fit with an R^2 of .94. Finally, we fit the Busemeyer's averaging model to predict such data as suggested by Kao and Wasserman (1993). In their development of it, they predict subjects' judgments as a linear function of the proportions of p_a, p_b, p_c, and p_d of the observations. For instance, $p_a = a/(a+b+c+d)$. Applied to our situation, this model becomes

$$50 + w_1 p_a - w_2 p_b - w_3 p_c + w_4 p_d$$

if we assume 50 as the indifference rating.[5] Fitting this model to the data, we obtained an R^2 of .93, with parameters $w_1 = 50$, $w_2 = 57$, $w_3 = 35$, and $w_4 = 7$. This is somewhat worse than the weighted ΔP model despite the fact that it involves estimating one extra parameter. Its relatively good fit (compared with the Bayesian model or the linear regression model) reflects the fact that it, too, predicts subject behavior in terms of the relative proportion of events rather than the absolute number of events. This, again, is evidence for a nearly total disregard of sample size.

We do not want to attach too much significance to these relative R^2s. All the R^2 values are relatively high because they capture the general sensitivity of subjects to the variables a, b, c, and d. The relative ordering of the R^2s across models just confirms what the qualitative analyses of the data indicated. The qualitative analyses indicated curvilinear effects of a through d and the disregard of sample size. Both of these effects are predicted by the ΔP model and by the averaging model.

Subject reports. We asked subjects what variables they were paying attention to. On the basis of their own verbal and written reports, the ways by which the ratings were generated can be grouped as follows: (a) the majority of the subjects (26 out of 40) said they had tried to take all four types of information into account. Six of these reported actually doing what the ΔP rules prescribe. That is, they tried to calculate and compare the proportion of side effects when given drug treatment and the proportion of side effects when given no drug treatment. Others were less precise about what they had done but said that they had paid attention to all four values, and many indicated sensitivity to ratios or proportions; (b) 9 subjects reported that they had ignored the cases in which the drug was not used and had attempted to compare the occurrence of side effects versus no side effects in the presence of drug treatment alone; (c) 5 subjects reported other strategies which we found hard to interpret.

Figure 4 presents the data separately for subjects in Groups a and b. As can be seen, true to their word, subjects in Group a were showing relatively large effects of c and d while subjects in Group b showed relatively weak effects. Mixing these two groups of subjects would produce a weighted ΔP rule. It is true that even the subjects in Group a were not giving as much weighting to c and d as they were to a and b. Perhaps, on some trials, they ignored the absence condition. Thus, the more general proposal is that subjects' data can be predicted as a mixture of responding to ΔP and to just $P_1 = a/(a+b)$.

These data contrast with the data reported by Wasserman et al. (1990). They asked subjects in advance about what cells they thought they should pay attention to. They found 32 subjects who said they should pay attention to all cells and 13 who said they should pay attention only to the presence cells. However, there was no difference among these subjects in the effects of the four cells. All subjects showed larger effects of the a and b cells than of the c and d cells. The most apparent difference between these two experiments is that the Wasserman et al. subjects gave their reports in advance of the experiment, whereas our subjects gave their reports after the experiment.

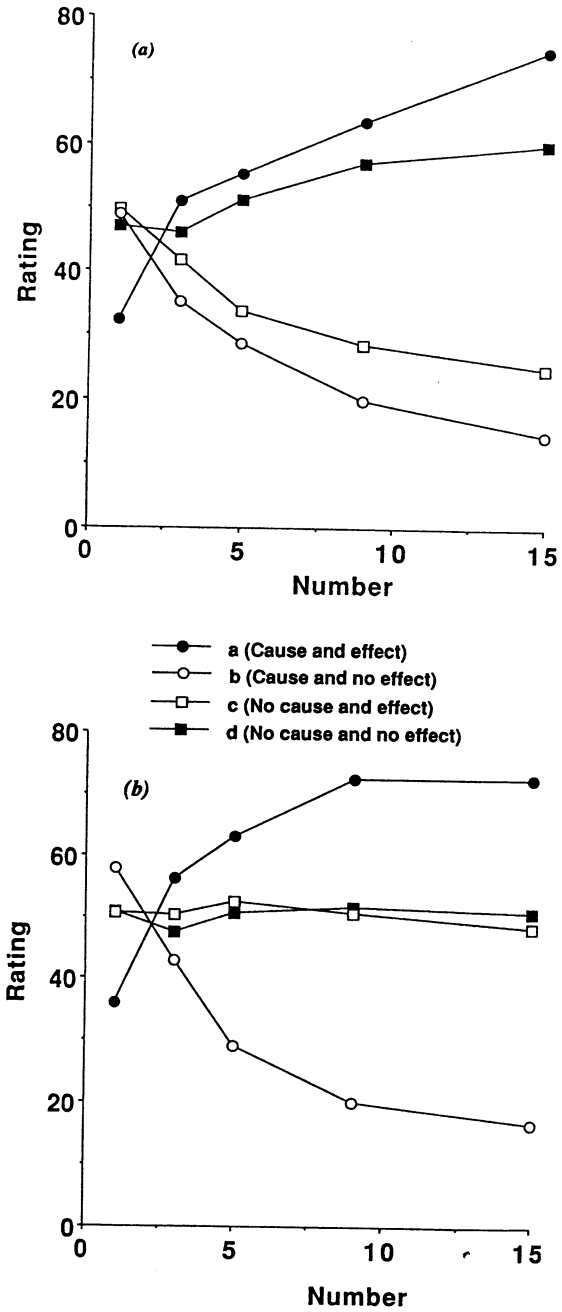

Figure 4. (a) Data from 26 subjects who reported paying attention to *a* through *d*. (b) Data from 9 subjects who reported paying attention to only *a* and *b*.

Discussion

Thus, it seems that the majority of the subjects are behaving according to the ΔP rule of Allan and Jenkins except that they weight the two proportions differentially (or perhaps they sometimes neglect $P_2 = c/(c-d)$). This model has recently been promoted by Wasserman (1990) and Wasserman et al. (1993), who have interpreted it in terms of the Rescorla–Wagner learning model (Rescorla & Wagner, 1972). Chapman and Robbins (1990) showed that the Rescorla–Wagner rule will produce an asymptotic level of association strength that corresponds to the value of ΔP. Wasserman et al. show that by assuming unequal learning rates in the various cells one can get a weighted ΔP model. Shanks (1987) has also promoted the interpretation of causal inference in terms of the Rescorla–Wagner model. He has also shown that the rate at which subjects approach asymptotic value can be predicted from the learning assumptions of that model.

The confirmation of the weighted delta rule in this experiment might seem to be evidence in favor of the interpretation of such judgments in terms of the Rescorla–Wagner model. However, subjects' reports certainly sounded very different from associative learning. They often claimed to be consciously comparing the two

probabilities. The fact that the data so cleanly divide on the basis of subject reports (Figure 4) should be taken as evidence that their reports are to be believed.

Allan and Jenkins (1980), Shaklee and Tucker (1980), and Wasserman et al. (1990) all have found that the behavior of a subset of subjects can be characterized as conforming to the ΔP rule. The subset of subjects graphed in Figure 4A would be those judged to correspond to the ΔP rule. One additional piece of information that we have from the present study is that subjects have conscious access to the rules that they are using and some even report explicitly calculating the ΔP quantity. Kao and Wasserman (in press) found that subjects were more likely to conform to the ΔP rule when the information was presented as summary numbers in a table rather than as individual cases (as in our experiment and most research). This would facilitate explicit calculations of proportions. By presenting summary totals in a table, the subject is relieved of having to count the four cells, which is a prerequisite to calculating the proportions.

Wasserman et al. (1993) reject the idea that subjects are explicitly calculating ΔP. They asked subjects to explicitly estimate the probability of an effect in the presence of the cause and, separately, the probability of the effect in the absence of the cause. They looked at how well these estimated conditional probabilities predicted subjects' causal ratings. They list two reasons for rejecting the ΔP model. First, the causal ratings correlated better with the objective conditional probabilities (r = .98) than they did with the subjects' estimates of conditional probabilities (r = .97). However, this is a very small difference. More impressive is the high value of the correlation with subjective probabilities. Perhaps noise in the estimations made them a poorer predictor than the objective probabilities. Their second reason for rejecting the ΔP model is that these subjects were more influenced by P_1 than by P_2. However, in our view, this difference in weighting is produced by subjects who always or sometimes fail to consider what happens in the absence of the cause. The weighted ΔP rule reflects this mixture of strategies. To address the Wasserman et al. arguments, we thought we should look at a critical difference between their paradigm and our own. The next three experiments will look at causal inference in the paradigm they used.

THE CONTINUOUS PARADIGM

Much of the recent evidence in favor of the ΔP model and its interpretation in terms of the Rescorla–Wagner learning rule has come from a paradigm that is different from the one we have considered so far. Rather than a discrete paradigm in which subjects are presented with explicitly paired causes and effects, the causes and effects occur continuously over time. This paradigm is more like the conditioning paradigms that gave rise to the Rescorla–Wagner model.

One can argue that the discrete paradigm represents a rather unnatural situation in that subjects are explicitly informed when causes and effects might go together. More normally, possible causes and effects do not appear in the environment so neatly paired. Rather, we encounter events continuously in time and we have to decide whether they are causally related. So, for instance, we might slam a door and observe a second later that a book falls off a shelf. We can entertain the hypothesis that the first event might have caused the second. The time between these two events is relevant to assessing whether or not there is a causal connection.

Wasserman (1990) reports typical research in a paradigm with human subjects that is designed to be analogous to the animal operant conditioning paradigms. Subjects were asked to explore whether tapping caused a light to come on. They were free to tap as often as they liked. The experiment was divided up into 1-sec intervals. Subjects chose to tap in about one-third of these intervals. Wasserman defined different conditions in terms of the probability that the light would be presented at the end of a 1-sec interval in which the subject tapped and the probability that the light would be presented at the end of a 1-sec interval in which the subject did not tap. In his Experiment 5, Wasserman created 25 conditions by factorially combining probabilities of .00, .25, .50, .75, and 1.00 of a light in a tap interval and a light in a nontap interval. In the case of a tap, these probabilities are referred to as $P(O|R)$ (for probability of outcome given response) and in absence of tap they are referred to as $P(O|\overline{R})$. These are analogous to the P_1 and P_2 that we defined for the discrete paradigm. Subjects had 60 sec to experiment with the causal relationship. They were asked to rate a causal relationship on a -100 to $+100$ scale, where -100 meant that the press prevented the light from turning on and $+100$ meant that the press caused the light to come on.

Wasserman (1990) noted in these data that $P(O|R)$ seemed to have a larger influence on subjects' judgments than did $P(O|\overline{R})$. He was successful in fitting a weighted ΔP rule to his data, just as we were with the data in Experiment 1. As we noted, Wasserman et al. (1993) argued that unequal weighting implied that subjects were not explicitly calculating $P(O|R) - P(O|\overline{R})$ as the ΔP rule would have. This is part of their reason for promoting the Rescorla–Wagner learning model with unequal learning rates. However, we have argued that these data could be interpreted as subjects's sometimes making their judgments on the basis of $P(O|R) - P(O|\overline{R})$ and sometimes on the basis of just $P(O|R)$, which would give an unequal weighting of the two proportions. We thought it would be worthwhile to collect some data in this continuous paradigm to see what subject reports were like and whether or not we could similarly split the subject population as we had in the discrete paradigm.

EXPERIMENT 2

The second experiment was an attempt to reproduce the standard continuous paradigm. Three values were used for $P(O|R)$ and $P(O|\overline{R})$—.167, .500, and .833. All

nine combinations of these probabilities were used. We were also interested in looking at effects of time and number of intervals: One condition involved thirty 750-msec intervals, one condition involved thirty 1,500-msec intervals, and one condition involved sixty 750-msec intervals. Comparison of these conditions would give us some sense of subject sensitivity to sample size. Combining these three types of problems with the different values of $P(O|R)$ and $P(O|\bar{R})$ yields 27 conditions. If a subject made a response in one of these intervals, the effect occurred at the end of the interval with probability $P(O|R)$, and if the subject's response did not occur in that interval, the effect occurred with probability $P(O|\bar{R})$. Wasserman (1990) looked at variables such as time of intervals and number of intervals and did not find any substantial effects. Shanks (1987) found effects of number of intervals when the number was very small. Shanks interpreted these early changes as reflecting the learning process by which the Rescorla–Wagner model reaches asymptote. The insensitivity to number of intervals beyond the first few might be seen as another case of subject insensitivity to sample size.

Method

Depending on the condition, subjects had either 22.5 or 45 sec in which to experiment with a flute. They were asked to judge whether clicking the mouse of the computer caused a flute icon on the monitor to play a tune. After experiencing one such problem, they were asked to judge its causal efficacy and then to judge another problem. Altogether, they judged 81 such problems in random order. These involved three replications of each of the 27 conditions. At the end of a problem, the subjects were asked to enter a number from −100 to 100 that best reflected their judgments of the causal relationship between the clicking and the icon. We used this response scale to correspond to the one used by Wasserman et al. (1993).

Subjects. Eighteen graduate and undergraduate students were recruited through an ad on the Carnegie Mellon University computer market bulletin board. They were each paid $8 for participating in the 1-h experiment. There were 10 male and 8 female subjects, of whom 11 had taken statistics courses. Neither sex nor the prior statistics course interacted with the results. Therefore, we will average the results over all subjects.

Apparatus. A Mac IIci computer was used to control the stimulus display and collect data from the subjects. The display was a 5×3 in. window at the center of an Apple 21-in. monochrome monitor.

Procedure and Instructions. Figure 5 shows the screen display. The subjects were given three practice trials at the beginning of the experiment and a 5-min rest halfway through the experiment. The following are the critical instructions:

> In this experiment your task is to find out whether clicking the mouse has any effect on whether or not the flute icon shown on the monitor sounds a tune. At any time you may choose to click the mouse or not click it. You can click it as often or as rarely as you like. However, to make a good judgment, you must pay attention to whether the flute plays the tune when you click the mouse and when you don't click it. The flute icon is highlighted when you click the mouse. Clicking the mouse has no effect whatsoever when the icon is still highlighted. Otherwise, you may click it at any time you like. Please release the mouse once a click is made.
>
> There will be 81 different problems, each lasting for either 22.5 or 45 seconds. In each problem the flute will sometimes sound when you click the mouse, and will sometimes sound of its own accord. The relationship between clicking the mouse and whether or not the flute sounds will be constant within each problem, but may well differ from one problem to the next. Your task is to choose an integer between −100 and 100 that best reflects the degree to which you believe clicking the mouse has an effect on sounding the flute icon. For example, choosing a negative number means you think clicking the mouse *prevents* the flute from playing the tune; while choosing a positive number means you think clicking the mouse *causes* the flute to produce sound. +100 indicates that clicking the mouse always causes the flute to sound, and −100 indicates that clicking the mouse always prevents the flute from sounding. Zero indicates that clicking the mouse has no effect on whether or not the flute sounds. Please type in the number at the prompt of the dialogue window.

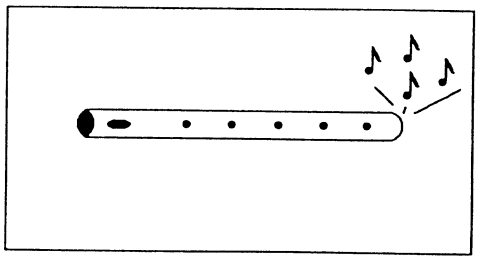

Figure 5. Example of stimuli from Experiment 1.

Results

The results of the three observations per condition were averaged for each subject and then subjected to a three-way analysis of variance. The analysis of variance reveals a significant effect of $P(O|R)$ and $P(O|\bar{R})$ [$F(2,34) = 158.1$, $MS_e = 1{,}111.3$; $F(2,34) = 127.7$, $MS_e = 712.8$, respectively] but no effect of the variable of time and number of events [$F(2,34) = .10$, $MS_e = 546.2$], substantially replicating Wasserman (1990). Figure 6 displays the effects of $P(O|R)$ and $P(O|\bar{R})$.[6] The weighted ΔP model was fit to the data and accounted for 97% of the variance. The regression equation is

$$\text{Rating} = -9 + 100P(O|R) - 72P(O|\bar{R}).$$

Thus, the data nicely replicate past results and are consistent, at this level of analysis, with what we obtained in the discrete paradigm. This helps establish that our procedures and subjects are not unlike those used by other researchers and so sets the context for other analyses.

When we interviewed the subjects, we found it quite impossible to interpret their reports as indicating that they had paid attention to *a*, *b*, *c*, and *d* events or some subset of these events. Unlike in the case with the discrete paradigm, they perceived no event boundaries. Thus, subjects could not easily calculate the number of *a*, *b*, *c*, and *d* events and no one reported trying to estimate either $P(O|R)$ or $P(O|\bar{R})$. Rather, they claimed to be responding to the rate at which the flute sounded when they were pressing and when they were not pressing the mouse. As one subject reported: "I clicked continuously for a few seconds, observed, then did not for a while and saw what made a difference in the beeping. The gaps between the beeps were most informative but I looked at the frequency as well." Another subject reported: "I would listen to a pattern and then try to alter it with the clicking. The more my clicking sped the pattern up—the higher the positive number—and vice

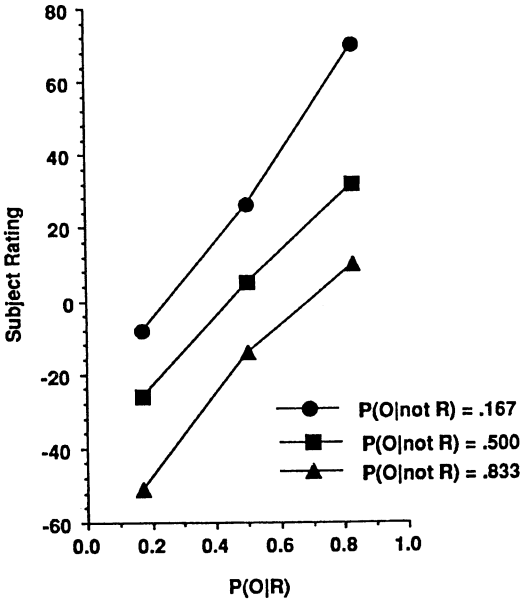

Figure 6. Results of Experiment 2. Effects of $P(O|R)$ and $P(O|\bar{R})$.

versa. Slowing the pattern down produced a negative number."

Thus, subjects claimed to be responding to the rate at which the beeps were occurring or the frequency of beeps in a self-defined click or no-click interval. They were not counting events of no beeps (corresponding to b and d). One implication of the reports is that subjects tend to cluster their clicking by clicking for a number of intervals and pausing for a number of intervals. This would suggest that subjects are quite likely to click in the current interval if they had clicked in the previous interval but not to do so if they had not. Indeed, this is what we observed with a .64 probability of clicking in the current interval if they had clicked in the previous interval, but only a .27 probability if they had not clicked in the previous interval.[7] By contrast, if they had been clicking uniformly throughout the experiment there would not have been such a difference.

It might seem strange that the weighted ΔP model does so well in predicting the data if subjects are responding to the rate of clicking and not to the probability. However, the probability manipulation controls the rate of clicking. It is true that we used two different event intervals (750 and 1,500 msec) and found no effect of this manipulation. However, the increase to 1,500-msec intervals would slow the rate of beeping in both click and no-click intervals and not affect the relative differences in rates. Thus, relative rate was totally confounded with relative probability. Wasserman and Neunaber (1986) and Shanks, Pearson, and Dickinson (1989) have found effects of the delay between response and event on causality judgments. In their experiments, they varied the interval between response and event but held constant the interval between events when subjects did not respond. Our third experiment was undertaken to see what would happen when we decorrelated relative rate and relative probability and had different delays in the case of a response and in the case of no response.

EXPERIMENT 3

In Experiment 3, the method, materials, and instructions to subjects were the same as those in Experiment 2. There were 15 conditions defined by the crossing of 5 contingency conditions with 3 time conditions. The contingency conditions were designed to create differences of $P(O|R) - P(O|\bar{R})$ of $-.67$, $-.33$, 0, .33, and .67. To create these differences, the following five pairs of values for $P(O|R)$ and $P(O|\bar{R})$ were used: .17 and .83, .17 and .5, .5 and .5, .83 and .5, .83 and .17. The time to the flute sounding (if it did sound) was $V+T$, where V was a random exponentially distributed variable with a mean of 150 msec and T took on different values depending on the timing condition. In one condition, the T was 450 msec when subjects clicked and 1,050 msec when they did not; in a second condition, T was 750 msec in both intervals; and in a third condition, it was 1,050 msec when the subjects clicked and 450 when they did not. If the subject clicked, the program always waited until the selected time before sounding the flute, irrespective of whether the subject clicked again or not. After an interval was up, the program would treat the next time as a no-click interval unless the subject clicked before the no-click time was up, in which case the program would switch at that point to timing a click interval. Each trial involved 48 intervals.

The purpose of this manipulation was to introduce a decorrelation between probability and rate. Table 3 illustrates this for the 15 conditions of the experiment. There we have the probabilities, $P(O|R)$ and $P(O|\bar{R})$, associated with each condition, the mean times, $T(O|R)$ and $T(O|\bar{R})$, of the intervals, and the rates defined as $R(O|R) = P(O|R)/T(O|R)$ and $R(O|\bar{R}) = P(O|\bar{R})/T(O|\bar{R})$. As can be seen, there are substantial variations in rates in conditions of constant differences in probabilities.

Method

Except for the different timing arrangement, the procedure in Experiment 3 was identical to that of Experiment 2. There were 60 problems which involved four replications of each of the 15 conditions. At the end of a problem, the subjects were asked to enter a number from -100 to 100 that best reflected their judgment of the causal relationship between the clicking and the icon.

Eighteen graduate and undergraduate students were recruited through an ad on the Carnegie Mellon University computer market bulletin board. They were each paid $8 for participating in the 1-h experiment. There were 12 male and 6 female subjects, of whom 12 had taken statistics courses.

Results

A within-subjects analysis of variance revealed significant effects of contingency condition [$F(4,68) = 85.85$, $p < .001$, $MS_e = 1{,}071$][8] and timing [$F(2,34) = 34.44$, $p < .001$, $MS_e = 656$] and a significant inter-

Table 3
Conditions of Experiment 3

Condition	$P(O\|R)$	$P(O\|\bar{R})$	$T(O\|R)$	$T(O\|\bar{R})$	$R(O\|R)$	$R(O\|\bar{R})$
1	.17	.83	.60	1.20	.28	.69
2	.17	.83	.90	.90	.19	.93
3	.17	.83	1.20	.60	.14	1.39
4	.17	.50	.60	1.20	.28	.42
5	.17	.50	.90	.90	.19	.56
6	.17	.50	1.20	.60	.14	.83
7	.50	.50	.60	1.20	.83	.42
8	.50	.50	.90	.90	.56	.56
9	.50	.50	1.20	.60	.42	.83
10	.83	.50	.60	1.20	1.39	.42
11	.83	.50	.90	.90	.93	.56
12	.83	.50	1.20	.60	.69	.83
13	.83	.17	.60	1.20	1.39	.14
14	.83	.17	.90	.90	.93	.19
15	.83	.17	1.20	.60	.69	.28

action between the two [$F(8,136) = 6.25$, $p < .001$, $MS_e = 284$]. The results are displayed in Figure 7. As can be seen, the subjects' data increase both as the contingency difference between click and no click increases and when the click interval is short relative to the no-click interval. The significant effect of timing means that a model that uses only probabilities, as does the ΔP model, is wrong. The effects of timing appear largest for intermediate values of $P(O|R) - P(O|\bar{R})$ where there are no floor or ceiling effects.

We compared a number of models for predicting subject data. One (the ΔP model) used the probabilities, and the other used the mean rate at which the flute was expected to sound per unit of time. These two factors were confounded in the previous experiment. Using probabilities, a regression equation that accounted for 84% of the variance was obtained:

$$\text{Rating} = -3 + 85P(O|R) - 78P(O|\bar{R}).$$

Using rates, the following equation, which accounted for 88% of the variance, was obtained:

$$\text{Rating} = -12 + 51R(O|R) - 35\,R(O|\bar{R}),$$

where $R(O|R)$ is rate of sound when clicking and $R(O|\bar{R})$ is rate when not clicking. Again, the difference in R^2 between the two models is not large because both capture much of the major trends in the data. The real evidence for the rate model is the significant effect of timing in the analysis of variance.

If we take seriously subject reports that they were responding to perception of relative rates of beeping, this suggests that we should be applying models from psychophysics as to how people compare perceptual quantities. One frequently used statistic is Weber's contrast, or the *grating contrast* (Barlow & Mollon, 1982), which is the difference between two magnitudes over their sum. In the context of our current situation, this quantity is

$$G = \frac{R(O|R) - R(O|\bar{R})}{R(O|R) + R(O|\bar{R})}.$$

The grating contrast has the advantage of being the right response scale for these experiments, varying from -1 to 1. To a good approximation, the grating contrast behaves like a logarithmic function of the ratio of rates. Correlating this quantity with the data accounts for 93% of the variance with one less parameter than either the ΔP or the ΔR models above. The best fitting equation is

$$\text{Rating} = 4 + 77G.$$

The grating contrast provides a better fit to the data because it makes the difference in rates relative to the absolute values of rates. One can imagine the subject detecting changes in the pattern over the base rate defined by the denominator of the grating contrast. A similar equation was fit to Experiment 1, and it accounted for 90% of the variance. This is less than the ΔP model's 97%, but the grating contrast model has one fewer parameter. The first experiment did not vary timing in a way that would produce discriminating results.

In this experiment, as in the previous one, subject reports were also quite explicit about the fact that they were responding to rates. One subject reported: "Listen first, try different patterns. Listen to intervals of silence and the flute. Both were important. Found mostly the time lags were important." Another reported: "I would wait 5 to 10 sec and listen to the rate of sounds when not clicking. Then I would click for about 5 sec and listen to the rate again. Then listen without clicking again."

Thus, as subjects report, it would seem that they were responding to the rates of sounds and that their judgments are best thought of as psychophysical judgments of the difference in the magnitudes of these rates.

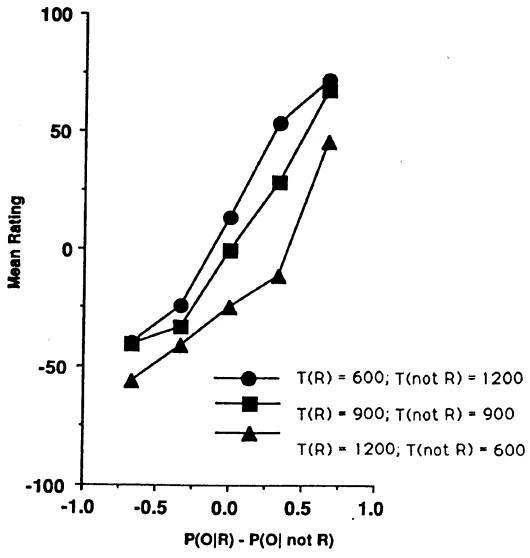

Figure 7. Results of Experiment 3. Effects of contingency and timing. Timing is given in terms of the constant time (in milliseconds). Another random time averaging 150 msec was added.

EXPERIMENT 4

Method

Having now shown that the critical variable seemed to be rate and not probability, we wanted to do a more parametric exploration of subject sensitivity to the time dimension. This experiment was like the third, except that $P(O|R)$ and $P(O|\overline{R})$ were 1—that is, every time interval ended with a sound. However, we manipulated the time orthogonally until the sound occurred in the presence and in the absence of a click. The mean waiting times when subjects did not click were 1,000, 2,000, and 4,000 msec, and the mean waiting times when subjects did click were 250, 500, 1,000, 2,000, 4,000, and 8,000 msec. Since the probabilities were 1, the rates in this experiment were simply 1 divided by the times. These times were randomly distributed according to an exponential distribution with the given means. Crossing these two variables resulted in 18 conditions. There were four replications of each condition in the experiment, for a total of 72 problems. Each trial lasted 40 sec and involved however many intervals would fit into that time period under the constraints of the condition, the pattern of clicking adopted by subject, and the random variability in interval generation. Otherwise, the procedure was identical to that of the previous experiments.

Eighteen graduate and undergraduate students were recruited through an ad on the Carnegie Mellon University computer market bulletin board. They were each paid $8 for participating in the 1-h experiment. There were 9 male and 9 female subjects, of whom 11 had taken statistics courses.

Results

Figure 8 shows the results of this experiment. There are main effects of click interval [$F(5,85) = 37.38$, $p < .001$, $MS_e = 1{,}688$], of no click interval [$F(2,34) = 18.79$, $p < .001$, $MS_e = 1{,}001$], and an interaction between the two variables [$F(10,170) = 2.93$, $p < .01$, $MS_e = 269$].[9] As the probabilities are held constant at 1 throughout the experiment, it is worth noting that a wide class of models are totally incapable of handling these data—including the Schustack and Sternberg (1981) linear model, the Bayesian model, any variation of the ΔP model (including the Rescorla–Wagner model), or Busemeyer's (1991) averaging model. One can only account for these data by a model that responds to rates of events.

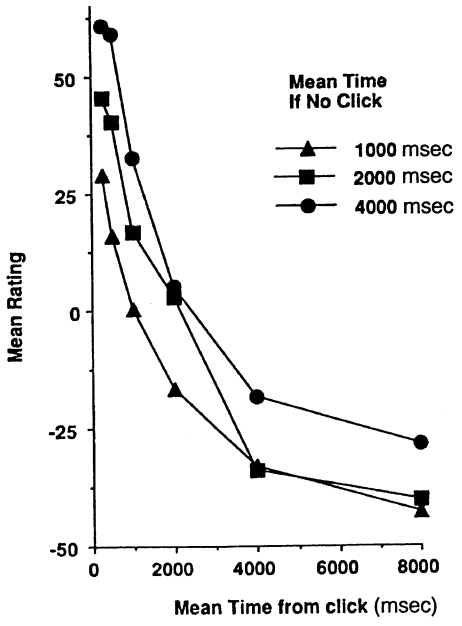

Figure 8. Results of Experiment 4. Effects of average times from click to sound and average times to sound, given no click.

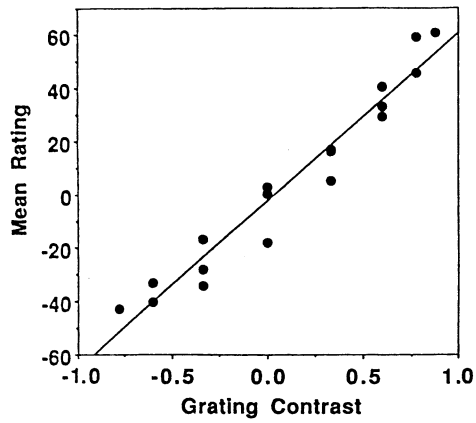

Figure 9. Results of Experiment 4 plotted as a function of the grating contrast.

We found the best fitting linear function to the rates to be

$$\text{Rating} = -1 + 20R(O|R) - 33R(O|\overline{R}),$$

which accounted for 76% of the variance. Again, a simple linear function of the grating contrast accounts for a great deal more variance (95%):

$$\text{Rating} = -3 + 63G.$$

This grating contrast model predicts a diminished effect of $R(O|\overline{R})$, given larger values of $R(O|R)$, which is in part the interaction displayed in Figure 8—a smaller effect of mean time if no click when there were larger values of mean time if click. Figure 9 illustrates the relationship between the grating contrast and subject ratings. It shows that different conditions with similar grating contrasts do produce similar ratings. Subjects were responding to the ratio of rates.

GENERAL DISCUSSION

The results of the first three experiments were relatively well fit by the weighted ΔP model. However, subject reports suggested rather different interpretations of these outcomes in the discrete and continuous paradigms. In the discrete paradigms, subject reports indicated that there was a mixture of responses on the basis of pure ΔP and responses on the basis of just P_1, with the resulting combination looking like a weighted ΔP. An analysis of the data sorted by subject report confirmed this indication. Subject reports in the continuous paradigm suggested that they were responding in terms of rates which tended to be confounded with probabilities.

This was confirmed in Experiment 3, which decorrelated rates and probabilities, and in Experiment 4, which held probabilities constant and varied rates.

Our research has found the discrete paradigm to be very different from the continuous paradigm, despite the tendency in the literature to treat them as the same. In the continuous paradigm, subjects do not respond to specific response-event combinations. Rather, they cut the period up into intervals both when they are responding and when they are not. Their judgments reflect the rate of the critical event in the response intervals versus the nonresponse intervals. This strategy reflects in part the fact that they can exercise control by responding in the continuous paradigm, which they cannot do in the discrete paradigm. However, more importantly, it reflects the fact that they cannot count pairings of responses and outcomes in the continuous paradigm. They do not know when the experimenter-defined intervals are up and whether or not the effect has occurred. Rather, they perceive a continuous stream of events. We have a classic case here of the difference that can occur between a subject's perception of an experiment and the experimenter's characterization of it.

It seems that the consideration that unites the two paradigms is that subjects respond in terms of whatever variables are perceptually salient and seem causally relevant. Subjects behave according to the ΔP model in the discrete paradigm because it is relatively easy to count events, which must be done in order to respond in terms of probabilities. As Kao and Wasserman (in press) have shown, subject behavior corresponds more to the ΔP model when the counts of events are provided for the subjects in summary tables.

Subjects do not behave according to the ΔP model in the continuous paradigm basically because it is not possible to count the various kinds of events. Rather, what subjects can do is to judge the perceptually salient feature, which is the rate of the event when they are responding and when they are not. It is significant that their judgments can be fit so well by a simple grating contrast, which also works well in similar psychophysical experiments. Gallistel (1990) has proposed a similar rate model in the conditioning literature as an alternative to the Rescorla–Wagner model which leads to a ΔP model prediction.

These results give us an interesting perspective both on the issue of the mechanisms for causal judgment and on the issue of the normative status of these mechanisms. It appears that subjects look for quantities that are both easy to compute and causally relevant, and that they make their judgments with respect to these. The more extreme such a quantity is in the direction expected by a causal relationship, the more evidence it provides for a causal relationship. However, subjects pay relatively little attention to the sample size on which this quantity is based. Thus, under a narrow definition, subjects cannot be viewed as making causal judgments in a normative manner. However, there are broader perspectives under which the issue is not so clear. For instance, Gallistel (personal communication, spring, 1994) has argued that it makes little sense, in foraging in a rapidly changing real world, to respond to long-term statistics and that organisms should respond to only their last few experiences.

There are two perspectives that one might take on these data. One is that subjects are engaging in causal inference but in a nonnormative manner. The second is that subjects are substituting a perceptual judgment for a causal judgment (in the narrow normative sense). One might seriously question whether subjects even know what it means to assess the probability that one event causes another. Certainly, psychologists and philosophers have had a hard time agreeing.

The implications of these results are essentially negative for theories that attempt to interpret causal inference as some sort of associative learning such as with recent uses of the Rescorla–Wagner model. In no real sense are subjects learning associations. The only "learning" involves estimating perceptual-like quantities, and often that involves explicit counting. These results are equally negative for such theories as Anderson's (1990) rational model or Cheng and Novick's (1992) focus-set ΔP model, which try to provide some rational reconstruction of what subjects are doing. We have already discussed problems that the rational model has. The focal-set ΔP model is very flexible and could deal with the mixture in the discrete paradigm simply by assuming that subjects have different focal sets—either all the events or just those where cause is present. However, it has no way of dealing with subjects' reliance on rates rather than probabilities in the continuous paradigm. Perhaps more fundamentally, these models cannot accommodate the fundamental assumption unifying our interpretation of discrete and continuous paradigms, which is that subjects are choosing to estimate whatever quantity is easy to calculate and also seems to have causal relevance. Our subjects are not judging probability of a causal relationship.

One could argue that the tasks that we asked our subjects to perform are artificial and that they have no relationship to causal inference outside the laboratory. While the requirement of assigning numbers on rather arbitrary scales adds an element of artificiality to the process, there are reasons to suspect that these judgments are related to judgments made outside the laboratory. The fact that the judgments are as regular as they are across conditions and as sensitive as they are to causally relevant variables is one argument that supports the suggestion that they are related to causal inference outside the laboratory. A second argument lies in the similarity of these results to research on conditioning and the similarity of those data to more natural data on foraging (Stephens & Krebs, 1986). Interestingly, research on foraging has also shown that animals seem to respond to rather simple easy-to-compute statistics that are correlated with optimal behavior (e.g., Kamil, Yoerg, & Clements, 1988).

Other evidence that these causality judgments are not epiphenomenal comes from the experiments of Chat-

losh, Neunaber, and Wasserman (1985) and Shanks and Dickinson (1989) in the continuous paradigm. They varied the delay of event and probability of the event and compared subjects who were asked to estimate causal force and subjects who were put in an instrumental situation and asked to maximize the number of events. They found that rate of responding in the instrumental condition mirrored the causal ratings given by subjects in the judgment condition. However, Chatlosh et al. did find that rate of responding increased as subjects had more experience with a positive ΔP (240 events vs. 60) while their ratings did not change. So, it is not entirely clear that the mechanism which controls their causal judgments is the same as the mechanism that controls their responding.

Another way some have questioned the artificiality of the task is to argue that such paradigms are basically devoid of any role for prior knowledge and so are quite unlike causal inference in most real-world tasks. While it is hard to judge the issue of relative frequency, there are cases in which people make causal inferences on the basis of contingency information alone. Many hypotheses in medicine start with simple correlations, and only much later is work done on the possible causal mechanisms grounded in prior biological knowledge. Much of what people learn about modern devices is based on raw contingency information—press this button and see what happens. The conditioning literature has shown that many organisms are capable of learning things on the basis of raw contingency. This is not to deny that prior knowledge, when it is available, can have a large effect on causal inference. However, it is a mistake to classify causal inference in the absence of prior knowledge as "artificial."

REFERENCES

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

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

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

ALLAN, L. G., & JENKINS, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning & Motivation*, **14**, 381-405.

ANDERSON, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

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

BARLOW, H. B., & MOLLON, J. D. (Eds.) (1982). *The senses*. Cambridge: Cambridge University Press.

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.

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

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

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

COHEN, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral & Brain Sciences*, **4**, 317-370.

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

FALES, E., & WASSERMAN, E. A. (1992). Causal knowledge: What can psychology teach philosophers? *Journal of Mind & Behavior*, **13**, 1-28.

GALLISTEL, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

HENLE, M. (1962). On the relation between logic and thinking. *Psychological Review*, **69**, 366-378.

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

JENKINS, H. M., & WARD, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, **79**(1, Whole No. 594).

KAHNEMAN, D., & TVERSKY, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430-454.

KAMIL, A. C., YOERG, S. I., & CLEMENTS, K. C. (1988). Rules to leave by: Patch departure in foraging blue jays. *Animal Behaviour*, **36**, 843-853.

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, & Cognition*, **19**, 1363-1386.

LIPE, M. G. (1982). *A cross-study analysis of covariation judgments*. Unpublished manuscript, University of Chicago, Center for Decision Research, Graduate School of Business.

MICHOTTE, A. (1963). *The perception of causality*. London: Methuen.

MILL, J. S. (1974). *A system of logic ratiocinative and inductive*. Toronto: University of Toronto Press. (Original work published 1843)

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

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

POPPER, K. R. (1972). *Objective knowledge*. Oxford: Oxford University Press, Clarendon Press.

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

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

SHAKLEE, H., & TUCKER, D. (1980). A rule analysis of judgments of covariation between events. *Memory & Cognition*, **8**, 459-467.

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

SHANKS, D. R. (1987). Acquisition functions in contingency judgment. *Learning & Motivation*, **18**, 147-166.

SHANKS, D. R., & DICKINSON, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition*, **19**, 353-361.

SHANKS, D. R., PEARSON, S. M., & DICKINSON, A. (1989). Temporal contiguity and the judgment of causality. *Quarterly Journal of Experimental Psychology*, **41B**, 139-159.

SHULTZ, T. R. (1982). Rules for causal attribution. *Monographs of the Society for Research in Child Development*, **47**(1, Serial No. 194).

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

STEPHENS, D. W., & KREBS, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

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

TOLMAN, E. C., & BRUNSWIK, E. (1935). The organism and the causal texture of the environment. *Psychological Review*, **42**, 43-77.

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

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: Academic Press.

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, & Cognition*, **16**, 509-521.

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

WASSERMAN, E. A., & NEUNABER, D. J. (1986). College students' responding to and rating of contingency relations: The role of temporal contiguity. *Journal of the Experimental Analysis of Behavior*, **46**, 15-35.

NOTES

1. The exception is Shanks (1985), who collected judgments through the course of presentation. His research will be discussed after Experiment 1.

2. Miriam Schustack has generously provided us with the raw data from this expeirment.

3. The ΔP model cannot be fit to their data because sometimes there were no observations to enable the calculation of P_1 or P_2—that is, either $a = b = 0$ or $c = d = 0$.

4. None of these effects interacted with whether the subjects had had a prior statistics course.

5. Because $\Sigma p_i = 1$, this model requires only four parameters, a, b, c, and d, and we do not lose anything by constraining the intercept to be 50.

6. As in Experiment 1, these effects did not interact with whether the subjects had had a prior statistics course.

7. On average, subjects clicked in 43% of the intervals, which is only a little higher than the rate Wasserman reports.

8. Again, none of these effects interacted with whether subjects had had a prior statistics course.

9. Again, there were no significant interactions with whether or not the subjects had had a prior statistics course.

(Manuscript received July 8, 1994;
revision accepted for publication August 1, 1994.)