Confounded: Causal Inference and the Requirement of Independence

Christian C. Luhmann (christian.luhmann@vanderbilt.edu)
Department of Psychology, 2 Hillhouse Ave
New Haven, CT 06511 USA

Abstract

One of the most important requirements for accurate causal inference is that there be no confounds; the cause being evaluated must occur independently of all other causes. When this requirement is not met, causal inferences are likely to be incorrect. The current study asks participants to judge how informative various situations are with respect to drawing causal inferences. Contrary to normative principles, participants appear to believe that many confounded situations are just as informative as unconfounded situations.

Keywords: Causal reasoning, inference, induction, confound

Introduction

To evaluate the causal influence of a given factor, experiments often use two groups: the experimental group that includes the factor of interest and the control group that does not. For example, a doctor who wants to test the efficacy of a new medication would give one group of patients the new medication and give another group of patients a placebo. The principles of experimental design mandate that these two groups ideally be identical in all respects except for the factor of interest (e.g. medication). If the two groups also differ in some other respect (e.g. gender), the experiment is said to be confounded.

Confounded situations make causal inferences difficult. Imagine that the results of our medication study demonstrated a significant outcome; the experimental group experienced fewer symptoms than the control group. Can we make conclusions about the effect of the new medication? The difference in symptoms may be because of the new medication. However, the difference in symptoms may also have been due to the unintended difference between groups (e.g. gender). How do we differentiate between these two alternatives? What is the correct inferential strategy in such situations?

Recognizing and dealing with the implications of confounds is part of life for seasoned experimentalists. However, lay reasoners must deal with similar quandaries when attempting to evaluate causes in everyday life. In fact, because people must typically rely on non-experimental observations, they are actually in a disadvantaged inferential position. It is currently unclear how and how well people deal with confounds in everyday reasoning. The current study aims to evaluate how confounded situations influence people's causal judgments.

Models of Causal Inference

In experimental design, confounds are typically described as an unintended difference between the experimental and control groups (as in the example above). More formally, a confound occurs when the cause of interest, C, and some alternative cause, A, occur in a statistically dependent manner (i.e., $P(A|C) \neq P(A|\sim C)$ where the tilde indicates absence)[1]. As will be shown, independence is a general requirement for valid causal inference.

Early models of causal inference made no special allowances for confounded situations. For example, ΔP (Jenkins & Ward, 1965) computes the causal influence of C on an effect, E, according to the following equation:

$$\Delta P = P(E|C) - P(E|\sim C) \qquad (1)$$

Thus, ΔP compares the probability of E in the presence C with the probability of E in the absence of C. When Equation 1 is significantly positive, C is judged to cause E.

The computation suggested by ΔP does not differentiate between confounded and unconfounded situations; Equation 1 is used in all situations. This is problematic because violations of independence can lead to erroneous causal inferences. For example, imagine that an alternative cause, A, has a significant influence on E and occurs more often when C is present than when it C is absent. P(E|C) will be large, not because of C's influence, but because of A's influence. The resulting ΔP will be significantly positive even when C has no influence. Thus, reasoners who use ΔP to evaluate causal influence can make erroneous causal inferences.

Fortunately, people's behavior does not conform to these predictions. For example, Spellman (1996) had participants observe an experiment in which two causes (a red and a blue liquid) were administered to various plants to see if they affected blooming. Participants received a set of observations, each of which described a specific plant. Each observation specified whether that plant received the red liquid, whether it received the blue liquid, and whether it bloomed. Some of the plant experiments were unconfounded; the two liquids were administered independently of each other. Other situations were confounded; independence was violated. Participants were then asked to evaluate the causal influence of one of the liquids.

When the situation preserved independence, participants' judgments matched the predictions of ΔP. When the

[1] Greenland, Robins, & Pearl (1999) refer to situations that violate independence as non-collapsible. For clarity, I will refer to such situations as confounds or violations of independence.

situation violated independence, participants' judgments deviated from ΔP. Participants appeared to recognize the confounded situations as such and adjust their reasoning accordingly. In these situations, participants' judgments suggest that they applied Equation 1, not to the entire set of data as ΔP suggests, but to a subset of the presented observations. Furthermore, participants did not simply use arbitrary subsets; the subsets used were those in which the two causes were independent of each other (generally using the observations in which the alternative cause was absent). That is, because the entire set of observations violated independence, participants used a subset of data in which independence was preserved. Spellman (1996) refers to this operation as conditionalizing and argues that such judgments are normative. In terms of avoiding problematic confounds, conditionalizing certainly appears to be normative since it establishes independence and thus permits valid causal inferences.

Thus, the suggestions of Spellman (1996) appear to render the problem of confounds moot. All that is needed is a focal set in which independence is satisfied and the problem of confounds disappears. However, imagine that our medication study is in fact confounded (e.g. more males in the experimental condition than in the control condition) but we have no record of our patients' genders. How do we construct a focal set in which medication-gender independence is guaranteed? It is unclear how one would guarantee independence in any given focal subset without information about the presence/absence of alternative causes. What are reasoners to do? Conditionalizing can't help when the alternative cause is unobserved because we cannot test for independence in any subset of observations. Because of this, even reasoners who wish to conditionalize cannot be guaranteed to make correct inferences.

This difficulty has lead researchers to begin describing the conditions required for normative (i.e., infallible) evaluation of causal influence. Inferential errors can then be evaluated with respect to these conditions and ultimate allow a more detailed understanding of people's behavior. For example, Cheng's (1997) power PC theory (PPC), relates ΔP to a normative quantity called causal power. When independence is violated, causal power cannot be accurately computed. Pearl (2000) derives a slightly different quantity (called PS) that also requires independence.

Given these analyses, what is the appropriate inference to make when confronted with unobserved alternative causes that violate independence? Though PPC and PS suggest the need for independence, neither prescribes measures to actually attain independence. Thus, one reasonable strategy might be to avoid making strong inferences at all. Given the uncertainty such situations create, it might be better to wait for additional information. If pressed to make a judgment, reasoners should certainly do so with little or no confidence.

Given people's sensitivity to observable violations of independence, perhaps they will be equally astute when faced with these more uncertain situations. What little work has been done on such situations (Perales, Cheng, & Catena, 2001, April; 2001, September) suggests that reasoners may be sensitive to unobserved confounds.

Participants in the Perales, et al. study were given information about how two causes varied with each other. The relationship between the two causes either preserved or violated independence. Participants were then given information about how the effect varied with one of the two causes (the target cause). The other cause was hidden from view, preventing participants from conditionalizing. In one condition, effect varied with the target cause perfectly (ΔP = 1.0). I refer to this as the Deterministic condition. In the other condition, the effect varied only moderately with the effect (ΔP = .66). I refer to this as the Probabilistic condition. Participants were then asked how informative the entire situation was with respect to evaluating the strength of the target cause. Participants rated unconfounded situations as significantly more informative than confounded situations. These findings suggest that reasoners are generally concerned with violations of independence and recognize the inferential dilemma they entail.

This study leaves many questions unanswered, however. For example, the Deterministic condition creates special circumstances for confounds. Contrary to the normative analyses, when the effect always and only follows the target cause, confounds do not generally prevent causal inferences. It is only when the target and alternative causes correlate perfectly that inferences are prevented (see below for a more detailed explanation). This suggests that participants' judgments in the Deterministic condition may not necessarily generalize to other confounded situations.

This realization points to a more general problem. The two causes in Perales' study confounded situation were perfectly correlated (i.e., $P(C|A)=1$, $P(C|\sim A)=0$). This condition provides a rather weak test of how confounds influence causal inferences. When the two causes are perfectly correlated, causal inferences about one cause (but not the other) are clearly impossible. It is unclear how reasoners might deal with less extreme violations of independence. It certainly seems possible that Perales's, et al. participants could have judged the confounded condition as uninformative because of the extreme nature of the violations used in the study. Violations of independence may be of less concern to people if the two causes sometimes occur separately. The case where causes are perfectly correlated may actually be a particularly salient confound and thus Perales' et al. findings may overestimate people's competence.

To clarify how violations of independence influence people's causal inferences, the current study uses a more fine-grained manipulation of independence. Instead of using only extreme violations of independence, the current study uses a spectrum of violations to better evaluate people's general ability to detect and deal with them.

Experiment

The current experiment consisted of three phases. In the first phase, participants were given information about how two causes, C and A, varied with each other. In the second phase, participants were given information about how the

	0.0 Condition		.25 Condition		.5 Condition		.75 Condition		1.0 Condition	
	A	~A	A	~A	A	~A	A	~A	A	~A
Covariation of the Two Causes C	8	8	10	6	12	4	14	2	16	0
~C	8	8	6	10	4	12	2	14	0	16
P(A|C)	.5		.625		.75		.875		1.0	
P(A|~C)	.5		.375		.25		.125		0.0	
P(A|C)-P(A|~C)	0.0		.25		.5		.75		1.0	

Figure 1 - The five conditions used in the current study. The frist row illustrates how the two causes, C and A, vary with each other. The next two rows contain conditional probabilities computed from the first row. The bottom row contains the difference between the two conditional probabilities

target cause, C, varied with the effect, E. In the third phase, participants were asked to make various judgments including the extent to which the information in the first two phases permitted causal inferences.

To systematically vary the independence of the two causes, the current study manipulates the difference between P(A|C) from P(A|~C) (varying from 0.0 to 1.0, see Figure 1). The only condition that preserves independence is the 0.0 condition where P(A|C) equals P(A|~C). All other conditions violate independence.

Like Perales, et al., the current study also manipulated the strength of the cause-effect relationship. In the Deterministic condition, the relationship between C and E was characterized by a ΔP of 1.0, C and E correlated perfectly. In the Probabilistic condition, the relationship between C and E was characterized by a ΔP of .75.

The Probabilistic condition provides a good test of how people deal with confounds because any violation of independence prevents valid causal inference. According to the normative analysis, the 0.0 condition should be rated as highly informative because it preserves independence between the causes. The other conditions violate independence and are thus uninformative according to the normative analysis (see Fig. 2).

Normatively, the Deterministic condition should show a different pattern. The 1.0 condition should be rated as uninformative because the two causes are inseparable. All other conditions should be rated as informative because, regardless of independence, they allow valid inferences (see Fig. 2).

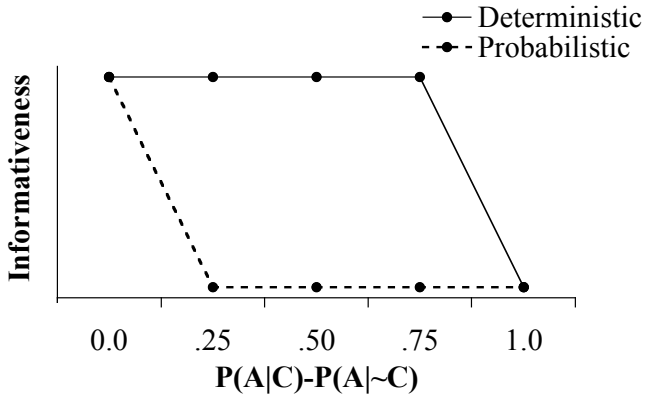

Figure 2 - Normative estimates of informativeness. In the Probabilistic condition, the 0.0 condition should be more informative than all other conditions. In the Deterministic Condition, the 1.0 condition should be less informative than all other conditions.

The Deterministic condition is unique because violations of independence do not generally prevent accurate causal inferences. It is possible to correctly disentangle the influence of the two causes. Take the Deterministic .5 condition. The alternative cause occurs in the absence of the target cause four times. On these four occasions the effect never occurs (the effect never occurs when the target cause is absent). This suggests that the alternative cause does not cause the effect. Conversely, the target cause and effect occur together on 16 occasions. On 12 of those occasions the alternative cause will also be present. To isolate the target cause, only the remaining four occasions must be used. Because all 16 occasions are identical (C and E both occur), it makes no difference which four occasions are used and which 12 are ignored. The target cause is always followed by the effect, suggesting that the presence of the target cause always causes the effect.

Figure 2 summarizes the normative pattern of informativeness for each condition. As can be seen, the patterns of informativeness are quite different for the Probabilistic and Deterministic conditions. In the Probabilistic condition, the 0.0 condition should be informative and all other conditions should uninformative. In the Deterministic condition, the 1.0 condition should be uninformative and all other conditions should be informative.

Of course, to mimic these normative patterns, people must be rather astute reasoners. They must first recognize how violations of independence impair causal reasoning. Furthermore, they must realize that the Deterministic condition presents a unique situation in which violations of independence do not always prevent valid conclusions.

In contrast, I would suggest that people might not recognize the inferential problems associated with "moderate" violations of independence (conditions, .25, .5, and .75). Furthermore, the analysis required to recognize the Deterministic condition as special may be too sophisticated for typical reasoners. Taken together, these possibilities suggest that the 1.0 condition should be consistently rated as less informative than other conditions.

Method

Fifty participants were told that they would be viewing the results from several drug trials. Each drug trial utilized two medications and recorded information about any side effects. The participants were told that they were going to be asked to evaluate the extent to which one of the drugs (the target drug) caused side effects. Half of the participants were assigned to the Deterministic condition and half were assigned to the Probabilistic condition. All participants received the five conditions illustrated in Figure 1 in pseudorandom order.

Each drug trial consisted of two sets of information presented on a computer. The first screen was intended to convey information about the correlation between two potential causes. This screen displayed 32 patients along with information about which of the two medications each received. No information about side effects was available at this point. How the two medications were distributed to the 32 patients is described in Figure 1, where A and C each denote a medication. After reviewing this information, participants moved on to the second screen. The second screen provided information about the target drug and the side effects for the same 32 patients. No information about the other drug was available at this point. Additionally, patients were shuffled between the first and second screen to prevent participants from simply matching the two sets of information. Participants were also given printed copies of all experimental information for reference.

After viewing these two screens of information, participants were asked a set of questions. To ensure that the independence manipulation was effective and noticeable, participants were asked to estimate P(A|C) as well as P(A|~C) in terms of frequency (e.g Sixteen patients received drug A, of those sixteen, how many also received drug B?). If these estimates were accurate, subsequent results cannot be attributed to participants' ignorance of independence. Most importantly, participants were asked to judge whether the information provided allowed them to judge the extent to which the target drug (by itself) caused the side effects (0-Definitely Not to 10-Definitley Yes); the same judgment elicited by Perales, et al. Participants were also separately asked to actually judge the strength of the target cause.

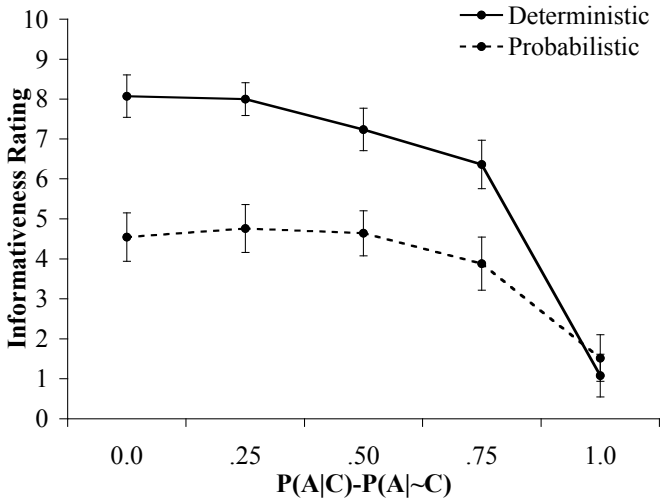

Figure 3 - Participants ratings of how informative each condition was. Error bars indicate standard error.

Empirical Results

A 2 (Deterministic vs. Probabilistic) by 5 (0.0, .25, .5, .75, 1.0) ANOVA was performed with repeated measures on the latter factor. The main effects of both factors were significant ($F(1, 47)=18.60$, $p<.0001$; $F(4, 188)=37.42$, $p<.0001$, respectively)[2] as was the interaction between them ($F(4, 188)=5.07$, $p<.001$). For clarity, the Deterministic and Probabilistic conditions will be further analyzed separately.

Before moving on, however, it is important to ensure that the violations of independence were noticeable to participants. To do so, participants' estimates of P(A|C) and P(A|~C) were compared to the actual conditional probabilities. Not surprisingly, frequency estimates were very accurate. No single estimate differed from the correct frequency by more than 1. Because of this, the correlation between mean estimates and true conditional frequencies was greater than .99, $p < .0001$. This finding allows for significantly cleaner interpretation of subsequent findings.

Deterministic Condition

Figure 3 presents participants judgments of informativeness. As can be see, judgments matched the normative predictions. When the two causes were perfectly correlated (the 1.0 condition), participants rated the situation as less informative than in the 0.0 condition ($t(24)=10.48$, $p < .0001$), the .25 condition ($t(23)=9.71$, $p < .0001$), the .5 condition ($t(24)=8.64$) , $p < .0001$, and the .75 condition ($t(24)=6.96$, $p < .0001$). In addition, the .75 was rated as significantly less informative than the 0.0 condition ($t(24)=2.34$, $p<.05$). No other significant differences were found.

Figure 4 displays participants' causal judgments of the target cause. Causal ratings of the target cause mirrored participants' informativeness ratings. The target cause was rated as weaker in the 1.0 condition than in any of the other conditions (all p's<.001). No other significant differences were found.

Probabilistic Condition

It is interesting to note that informativeness ratings were lower for the Probabilistic condition than for the Deterministic condition. This suggests that participants believe probabilistic situations to be inherently less informative than deterministic situations. Such a belief is not predicted by normative analyses. According to these accounts, probabilistic data is just as informative as deterministic data. Participants, on the other hand, may believe that causes are naturally deterministic and that apparently probabilistic relationships result from "noise."

The pattern of participants' informativenss judgments in the Probabilistic condition also did not match the normative predictions. Instead, participants' judgments were similar to

[2] Some participants chose not to respond to some of the queries. The variable degrees of freedom in the following analyses reflect this fact.

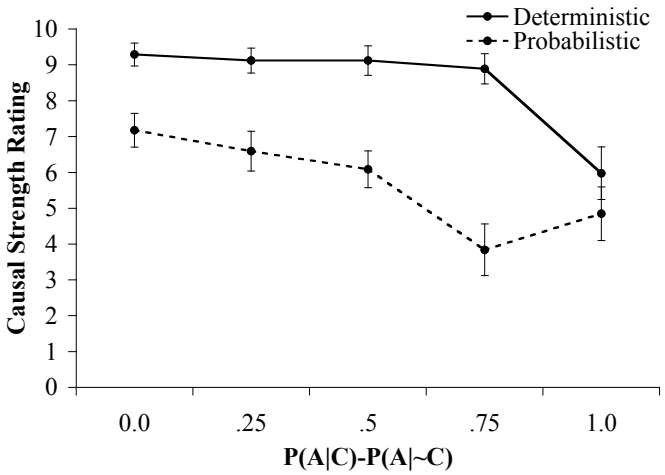

Figure 4 - Participants' causal strength ratings. Causal strength could range from 0 to 10. Error bars indicate standard error.

those in the Deterministic condition though lower overall (see Figure 3). As in the Deterministic condition, preservation of independence (the 0.0 condition) lead to participants to judge the situation as significantly more informative than when the two causes were perfectly correlated (M=4.54 and 1.52 respectively, $t(24)=3.71$, $p<.005$). Contrary to the normative prediction, the remaining conditions (.25, .5, and .75) were all rated as significantly more informative than the 1.0 condition ($t(24)=4.29$, $t(24)=3.65$, and $t(24)=3.51$ respectively; all p's $< .01$). No other significant differences were found.

Participants' causal judgments are displayed in Figure 4. The overall pattern is somewhat different than that of the Deterministic condition. The perceived strength of the target cause declined steadily as P(A|C) and P(A|~C) diverged, reaching a minimum in the .75 condition. The target cause in the .5 condition was judged to be marginally weaker than the target cause in the 0.0 condition ($t(23)=1.71$, $p=.1$). The target cause in the .75 condition was rated as weaker than the target cause in the .5 condition ($t(23)=3.24$, $p<.01$).

Summary When the two causes were perfectly correlated (the 1.0 condition), participants rated the situation as uninformative. Participants rated all other situations as significantly more informative. This was the case in both the Deterministic condition (where such judgments are accurate) as well as in the Probabilistic condition (where such judgments are inaccurate). Instead of believing that violations of independence prohibit causal inferences, participants appeared to believe that only complete overlap between the two causes prevented inferences.

Analytical Results

Participants' inappropriate beliefs about the informativeness of confounded situations could obviously lead to erroneous causal inferences. However it would be more compelling to actually assess accuracy instead of contemplating hypothetical possibilities. How appropriate were participants' causal strength judgments?

To help answer this question, I computed the maximum likelihood estimate (MLE) of the strength of both the target and alternative causes. Unlike other measures of causal influence (e.g. Cheng, 1997; Pearl, 2000), the MLE does not compute infallible estimates of causal influence. Instead, the MLE provides the "best guess" estimate of causal influence, which may turn out to be incorrect.

The current design withholds two important quantities from participants (i.e., the causal strength of the target cause and the causal strength of the alternative cause). It is possible to compute the likelihood with which each possible combination of values for these two quantities would produce the observed data. The most likely pair is then deemed to be the MLE. For this analysis, I assume that there are only two generative causes (the target and alternative cause) and that they combine their influence in accordance with the rules of a noisy-or gate (Glymour, 1998). The resulting parameters characterize the degree to which each cause is sufficient to bring about the effect (similar in this respect to Cheng, 1997 and Pearl, 2000).

The estimated strengths for the target cause are shown in Figure 5. As discussed above, the Deterministic condition allows for valid inferences. Thus, it is not surprising that the estimates match people's causal judgments quite well in the Deterministic condition. The estimated causal strength of the target cause was maximal (1.0) in all but the 1.0 condition. The 1.0 condition creates an ambiguity. The observed data could have result when *either* of the two causes is maximally sufficient, regardless of the other cause's strength. This ambiguity may explain why people believed the 1.0 condition to be highly uninformative and the associated drop in their causal judgment.

Causal judgments in the Probabilistic condition were also similar to the MLE. The computed estimates of the target cause's strength steadily decreased as P(A|C) and P(A|~C) diverged, eventually reaching a minimum of zero in the .75 condition. Participants' causal judgments demonstrated approximately the same pattern. Judgments in the .5

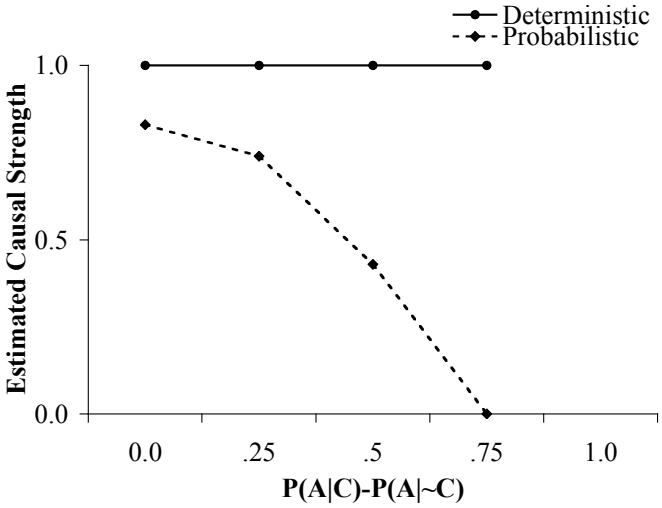

Figure 5 - MLE for strength of the target cause. Estimates of causal strength could range from 0 to 1.

condition were less than those in the 0.0 condition and judgments in the .75 condition were lower than those in the .5 condition. It is unclear whether the causal judgments in the .75 condition can be treated as minimal (M=3.84), though nearly a third of participants gave a causal rating of zero, making it the modal response. The MLE is again undefined in the 1.0 condition, this time because no combination of causal strengths will result in the observed data.

Summary In the Deterministic condition, participants' causal judgments were highly accurate. This is probably due to the fact that confounding in this condition does not cause the typical difficulties. In the Probabilistic condition, causal judgments were reasonable though fallible. Indeed, as the difference between P(A|C) and P(A|~C) increases in the Probabilistic condition, the likelihood of the MLE actually being correct diminishes rapidly. Nonetheless, people's judgments were quite close to the "best guess" values of the MLE.

Discussion

The results of the current study suggest that confounds do not necessarily have the impact on people's reasoning that they should. Confounds (i.e., violations of independence) generally limit the ability to make valid causal inferences. However, when asked to rate how informative various experimental situations were, participants' ratings appeared to be insensitive to the presence of confounds. This ignorance, could certainly lead to erroneous causal inferences in a variety of everyday situations.

Despite their erroneous judgments of informativeness, participants appear to have made rather reasonable causal strength judgments. This finding, however, does not diminish the error people make by ignoring violations of independence. Though the MLE provides the single most probable estimate, it is highly likely to be incorrect in the confounded situations (at least in the Probabilistic condition). An ideal reasoner would be able to recognize confounds as inferential obstacles *as well as* make reasonable causal strength judgments. Doing so would allow the good causal judgments without unwarranted confidence.

Why did participants incorrectly believe the confounded situations to be so informative? It is plausible that the full impact of confounded situations is not a completely integrated part of people's causal reasoning repertoire. As reviewed above, people do seem to control for alternative causes when they are observable. However, such findings may overestimate people's competence because the presence of a salient alternative cause may act as a reminder of the potential ambiguity interpretation must face. In everyday situations, where situations are assuredly messier, reminders of the inferential danger may not be as obvious.

In addition, when faced with an effect that clearly varies with a salient potential cause (as in the current study), people may often be tempted to jump to causal conclusions without consulting alternative explanations. Such a lack of critical analysis would explain a variety of erroneous beliefs. Superstitions are one obvious example. People often hold beliefs about the causal influence of obviously irrelevant actions (e.g. wearing a particular article of clothing causing a favorite sports team to win). These same people may also be able to acknowledge that alternative explanations exist for the desired outcome when pushed to do so. Similarly, I would speculate that at least some of the participants in the current study would acknowledge the uncertainty inherent in the experiment if it were pointed out to them. This discrepancy may suggest tension between two modes of reasoning: an automatic mode that is stimulated by the compelling cause-effect covariation and a more deliberate mode that may be able to resist such temptation (see Kahneman, 2003).

Acknowledgments

I would like to thank Woo-kyoung Ahn and Jessecae Marsh for comments on drafts of this paper.

References

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

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines,* 8, 39-60.

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science, 14*, 29-46.

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

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697-720.

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

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York, NY: Cambridge University Press.

Perales, J. C., Cheng, P. W., & Catena, A. (2001, April). Reasoning as scientists: Intuitive analogues of control procedures in experimental design. Associative Learning Symposium. Gregynog, Wales.

Perales, J. C., Cheng, P. W. & Catena, A. (2001, September). How do people control extraneous variables in causal learning tasks? Twelfth International Congress of the Spanish Society of Comparative Psychology. San Sebastián.

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