Journal of Experimental Psychology:
Learning, Memory, and Cognition
2011, Vol. 37, No. 3, 568–587

© 2011 American Psychological Association
0278-7393/11/$12.00 DOI: 10.1037/a0022970

Expectations and Interpretations During Causal Learning

Christian C. Luhmann
Stony Brook University

Woo-kyoung Ahn
Yale University

In existing models of causal induction, 4 types of covariation information (i.e., presence/absence of an event followed by presence/absence of another event) always exert identical influences on causal strength judgments (e.g., joint presence of events always suggests a generative causal relationship). In contrast, we suggest that, due to expectations developed during causal learning, learners give varied interpretations to covariation information as it is encountered and that these interpretations influence the resulting causal beliefs. In Experiments 1A–1C, participants' interpretations of observations during a causal learning task were dynamic, expectation based, and, furthermore, strongly tied to subsequent causal judgments. Experiment 2 demonstrated that adding trials of joint absence or joint presence of events, whose roles have been traditionally interpreted as increasing causal strengths, could result in decreased overall causal judgments and that adding trials where one event occurs in the absence of another, whose roles have been traditionally interpreted as decreasing causal strengths, could result in increased overall causal judgments. We discuss implications for traditional models of causal learning and how a more top-down approach (e.g., Bayesian) would be more compatible with the current findings.

Keywords: learning, causal reasoning, Bayesian inference, recency, primacy

When evaluating causal relationships, the covariation between events (see Figure 1) is one of the most crucial cues. A number of models (Busemeyer, 1991; Cheng, 1997; Einhorn & Hogarth, 1986; Jenkins & Ward, 1965; Rescorla & Wagner, 1972; Schustack & Sternberg, 1981; White, 2002) have been proposed, each specifying how to transform covariation into causal judgments. In many of these models, events of a given type (e.g., the joint presence of events or Cell A in Figure 1) always exert identical influences on judgments. In this article, we question this implementation and argue that, while learning about novel causal relationships, people make dynamic interpretations of events, which, in turn, result in dynamic influences of events on causal strength judgments.

In this introduction, we describe two sets of classic causal induction models to illustrate uniform, static influences of covariation information. We then discuss why covariation information may play a more dynamic role and how such flexibility could explain order effects in causal learning. We also briefly explain how our account is more compatible with the recent shift to a more top-down approach to causal learning.

Static Influence of Covariation

One important set of causal induction models is rule based, based on ΔP (Jenkins & Ward, 1965), which is

$$\Delta P = \left(\frac{A}{A + B}\right) - \left(\frac{C}{C + D}\right), \tag{1}$$

where each letter (A, B, C, and D) represents the frequency of observations from the corresponding cell of Figure 1. Positive ΔP values indicate a generative causal relationship (Event C produces Event E), and negative ΔP values indicate a preventative causal relationship (Event C prevents Event E). From Equation 1, one may readily see how each of the four types of observations should influence causal judgments. Observations from Cells A and D increase ΔP, holding all other cells constant. Observations from Cells B and C decrease ΔP, holding all other cells constant.

Another rule-based model, the power PC theory (Cheng, 1997), also suggests similar roles for each category of observation. When ΔP is positive, the generative causal power is computed as follows, and just as in ΔP, observations from Cell A increase the generative causal strength, and observations from Cell B decrease the generative causal strength.[1]

$$\textit{generative power} = \frac{\Delta P}{1 - \left(\frac{C}{C + D}\right)} \tag{2}$$

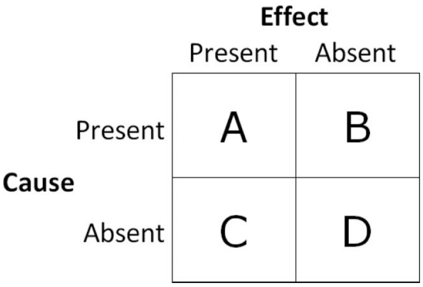
When ΔP is negative, the preventative causal power is computed as follows. Observations from Cell A decrease the preventative causal strength (i.e., making it more positive), and observations from Cell B increase the preventative causal strength (i.e., making it more negative).

Christian C. Luhmann, Department of Psychology, Stony Brook University; Woo-kyoung Ahn, Department of Psychology, Yale University.

This project was supported by National Institute of Mental Health Grant R01 MH57737 to Woo-kyoung Ahn. We thank Helen Randazzo, Evelyn Starosta, and Marla Krukowski for help with data collection and members of the Ahn lab for helpful discussion and feedback.

Correspondence concerning this article should be addressed to Christian C. Luhmann, Department of Psychology, Stony Brook University, Stony Brook, NY 11794-25001. E-mail: christian.luhmann@stonybrook.edu

[1] Under certain boundary conditions, these generalizations will not hold. For example, when ΔP is positive and $P(E|{\sim}C) = 1$, observations from Cells A or B will have no influence because causal power cannot be computed.



		Effect	
		Present	Absent
Cause	Present	A	B
	Absent	C	D

Figure 1. A contingency table summarizing the covariation between two binary events. Each cell of the table represents one possible observation.

$$preventative\ power = \frac{\Delta P}{\left(\frac{C}{C+D}\right)} \quad (3)$$

A second set of classic models is associative, such as the Rescorla–Wagner model (RW hereafter; Rescorla & Wagner, 1972). When learning about a single cause, RW updates the association between events on each trial according to

$$\Delta V = \alpha\beta(\lambda - \Sigma V), \quad (4)$$

where ΔV represents the change in the association between the cue and outcome resulting from the current trial and α and β represent learning rate parameters for the cue and outcome, respectively. The parameter λ takes on a value of 0 when the outcome is absent and is typically assumed to take on a value of 1 when the outcome is present. The quantity ΣV represents the sum of the associative strengths of all cues present on the current trial.

In nearly all situations, ΣV, the summed associative strength, will fall between 0 and λ (see Appendix A). Thus, when encountering an observation from Cell A, $(\lambda - \Sigma V)$ will be positive, resulting in a positive ΔV and increasing the strength of the association between events. When encountering an observation from Cell B, $(\lambda - \Sigma V)$ will be negative because λ is 0, resulting in a negative ΔV and decreasing the strength of the association between events. RW does not update the strength of causes when they are absent, so observations from Cells C and D do not alter the strength of causes (cf. Van Hamme & Wasserman, 1994). Thus, RW increases causal strength when encountering observations from Cell A and decreases strength when encountering observations from Cell B, much like the rule-based theories reviewed above.[2]

Dynamic Influences of Covariation Information

The classic models discussed so far are heavily data driven, imposing static roles on covariation information. However, there has been a recent shift within the field of causal learning that emphasizes the idea that existing beliefs or theories can shape learning from covariation data. For example, there is evidence that the use of covariation information may be influenced by learners' beliefs about causal structure (Griffiths & Tenenbaum, 2005; Waldmann, 1996; White, 1995) and by beliefs about how multiple causes interact to produce their effects (Beckers, De Houwer, Pineno, & Miller, 2005; Lucas & Griffiths, 2010; Vandorpe, De Houwer, & Beckers, 2007). To account for the joint influence of prior beliefs and covariation information, there have been several promising proposals (Griffiths & Tenenbaum, 2005; Lu, Rojas, Beckers, & Yuille, 2008; Lucas & Griffiths, 2010) formulated utilizing the framework of Bayesian inference. As is more thoroughly discussed in the General Discussion, these previous proposals were developed to account for the role of complex causal beliefs. While consistent in spirit, our current proposal pertains to the influence of much simpler causal beliefs (e.g., current estimates of causal strength) in much simpler causal contexts (e.g., a single cause and a single effect), which are not immediately explained by these models.

Specifically, we propose that the direction in which individual pieces of covariation information can sway causal strength judgments is much more dynamic than implemented in many of the classic models of causal learning. This flexibility stems from the fact that any single piece of covariation information (e.g., an observation from Cell A) can be interpreted so as to be consistent with a variety of causal beliefs (i.e., generative, preventative, or no relationship at all).

Table 1 illustrates such dynamic interpretations. To be concrete, imagine that one is tracking the covariation between a new medication (the ostensible cause) and pain (the ostensible effect), and one observes a patient who took the medication and experienced pain (i.e., Cell A). The patient, like the learning models discussed in the previous section, might lean toward interpreting the medication as causing the pain (i.e., generative hypothesis). Yet, if a pharmaceutical company manufactured this medication to control blood pressure, it would argue that the medication had nothing to do with the pain (i.e., a neutral interpretation) and that something other than its medication must have instead caused the pain. Alternatively, if a pharmaceutical company actually manufactured this new medication to alleviate pain, it might argue that the medication is generally effective in alleviating pain (i.e., negative hypothesis) but merely failed to do so on this occasion because something went wrong (e.g., a drug interaction, etc.). In this way, any given observation can be interpreted so as to be consistent with either a positive, neutral, or negative causal hypothesis.[3]

We further argue that evidence cannot only be dynamically interpreted but also must be dynamically used to modify one's current causal beliefs. For example, if an observation from Cell A is given a negative interpretation, this supposedly positive information could produce negative changes in learners' causal beliefs. The traditional view, as discussed in the previous section, has been that Cells A and D always result in more positive beliefs and that Cells B and C always result in more negative causal beliefs. The

[2] Unlike the rule-based models, however, RW predicts that the influence of a given observation depends, in part, on how recently that observation was encountered; more recent experience is more influential than less recent. Nonetheless, as we discuss below, RW assumes that the direction of influence (i.e., whether to increase or decrease the associative strength) is invariant; Cell A increases associative strength, and Cell B decreases associative strength.

[3] In practice, some interpretations may more easily serve as a default for a given observation (e.g., positive causal interpretation for Cell A), possibly because they prefer simpler explanations (Lombrozo, 2007) and making nontraditional interpretations (e.g., negative interpretation for Cell A) requires postulating an alternative cause or preconditions that must be met, which is more complex and may require additional cognitive effort.

Table 1
The Range of Causal Explanations for Individual Pieces of Covariation Information

	Possible interpretation		
Cell	Positive (generative)	Neutral (no influence)	Negative (preventative)
A (CE)	C produced E.	It just so happened that E followed C.	C failed to prevent E because something went wrong.
B ($C\bar{E}$)	C failed to produce E because something went wrong.	It just so happened that the absence of E followed C.	C prevented E.
C ($\bar{C}E$)	Something other than C produced E.	It just so happened that E followed the absence of C.	E happened because C was absent and something other than C produced E.
D ($\bar{C}\bar{E}$)	E did not happen because C was absent.	It just so happened that the absence of E followed the absence of C.	Nothing caused E.

contrast between these views is an unexplored dimension of causal learning and is the main focus of the current study.

Order Effects

The paradigm we use to demonstrate dynamic interpretations and dynamic use of covariation information involves manipulating the order in which observations are presented so that learners acquire different initial expectations, which, we argue, would result in different interpretations for identical observation later in the sequence and in different causal strengths. Indeed, several researchers have reported systematic effects of presentation order on causal judgments. Here, we briefly review this literature.

For instance, in López, Shanks, Alamaraz, and Fernández (1998), participants observed 160 trials in which they learned about diseases and symptoms. In the strong–weak condition, the first half of the sequence suggested that a symptom was strongly associated the disease, and the second half of the sequence suggested that they were weakly associated. In the weak–strong condition, the order of the two blocks was reversed. The final causal strength ratings were significantly higher for the weak–strong condition than for the strong–weak condition, demonstrating recency effects.

In contrast, several researchers have demonstrated primacy effects, in which early experience is more influential than recent experience (Chapman & Chapman, 1969; Dennis & Ahn, 2001; Yates & Curley, 1986). For instance, Dennis and Ahn (2001), from whom our Experiment 1 is derived, developed a positive block in which the bulk of trials consisted of Cells A and D (see Figure 1) and a negative block in which the bulk of trials consisted of Cells B and C. Participants who observed the positive block followed by the negative block gave overall causal strength ratings that were higher than those who observed the negative block followed by the positive block.

Our expectation-based account, described earlier, can readily explain the primacy effect. Learners would initially develop some hypothesis about how events are related to each other. The initial hypothesis would then alter how covariation information is interpreted later in the sequence. Since the initial hypothesis is developed based on data presented early in the sequence, earlier observations have more influence in overall causal strength judgments than later observations.

This expectation-based account is also compatible with the recency effects (López et al., 1998). Marsh and Ahn (2006) noted that López et al. (1998) had learners simultaneously learn four different sets of stimuli (each representing a separate condition) that were intermixed into a single sequence and argued that this could have prevented learners from forming any initial expectation that would have otherwise led to a primacy effect.[4] Indeed, Marsh and Ahn found a primacy effect in a simplified version of the López et al. study. Furthermore, they demonstrated a significant correlation between verbal working memory capacity and order effects during learning such that those learners with larger verbal working memory capacities exhibited greater primacy effects. Marsh and Ahn suggested that this latter result was due to those with greater working memory capacities being better able to maintain their hypotheses and/or utilize their prior expectations to modulate information processing. These findings support the idea that, when learners are able to do so, they develop hypotheses early in learning and that these hypotheses influence the processing of subsequent experiences.

Although the expectancy-based proposal is consistent with the empirical findings on order effects so far, there has been no direct empirical demonstration that the order effects are indeed related to (or caused by) dynamic interpretations. The order effects found in previous studies (Dennis & Ahn, 2001; López et al., 1998; Marsh & Ahn, 2006) could have been generated by noninterpretational mechanisms (Danks & Schwartz, 2005), such as increased or decreased attention (e.g., fatigue, boredom, or context change; Anderson & Hubert, 1963; Hendrick & Costantini, 1970; Stewart, 1965), or by discounting of later information as being less reliable or valid than earlier information (Anderson & Jacobson, 1965).

Overview of Experiments

The current set of experiments was designed to directly test whether learners flexibly interpret covariation information during learning and whether such interpretations actually affect causal learning. We took two separate approaches for this purpose.

In Experiment 1, learners were directly asked to interpret individual observations during the course of an otherwise traditional learning paradigm. We predicted that learners' interpretations would substantially deviate from the traditional roles to be more consistent with their existing hypotheses.

[4] Alternatively, others (Collins & Shanks, 2002; Matute, Vegas, & De Marez, 2002) have suggested that the recency versus primacy effects depend on how frequently learners are asked to evaluate the relationships being learned.

We further anticipated that the learners' interpretations would be related to their causal strength judgments regardless of whether learners exhibited primacy or recency or no order effects at all. That is, we were not interested in whether primacy or recency effects better describe causal learning or under what circumstances primacy or recency effects occur. Instead, our strategy was to take advantage of the fact that both primacy and recency effects can occur during learning (e.g., by way of working memory load; Marsh & Ahn, 2006) and to use these effects to highlight the interpretational flexibility of covariation data and how they relate to causal strength judgments as expressed in terms of recency or primacy effects.

Experiment 2 attempted to go a step beyond simply relating interpretations and learning. Instead, we sought to provide definitive evidence that covariation information can exert influences that directly oppose traditionally assumed roles.[5] That is, we attempted to demonstrate that adding ostensibly negative covariation information (i.e., Cells B and C in Figure 1) to a trial sequence could lead to more positive causal beliefs and that adding ostensibly positive covariation information (i.e., Cells A and D in Figure 1) could lead to more negative causal beliefs. Such a demonstration would be the first of its kind and would be beyond the scope of nearly all currently implemented models of causal learning. Thus, it would provide particularly important insight into the role of interpretational flexibility in causal learning.

Experiment 1

In Experiment 1, participants observed a series of event pairs and made a causal strength judgment at the end of the sequence (e.g., Luhmann & Ahn, 2007; Shanks, Holyoak, & Medin, 1996; Spellman, 1996). During learning, participants were occasionally prompted to select an interpretation of a trial that they had just observed. The orders of trials were manipulated to be either positive–negative or negative–positive as in Dennis and Ahn (2001). Experiment 1A used this standard paradigm, whereas Experiments 1B and 1C introduced additional experimental manipulations to further explore learners' dynamic interpretations.

Experiment 1A

Method

Participants. Thirty-two Yale University (New Haven, CT) undergraduates participated for partial course credit.

Materials and procedure. Stimuli (see Figure 2 for an example) consisted of novel medications (e.g., DJE-143) and physical symptoms (e.g., increases in blood pressure or increases in weight). Participants were told to determine what influence the medications had on these symptoms. Each participant learned about two medication–symptom pairs across two separate experimental conditions (see below). In each condition, participants progressed through a sequence of trials, each describing a different patient. For each patient, participants were told (a) whether that patient took the medication and (b) whether that patient developed the symptom. Progress through the sequence was self-paced.

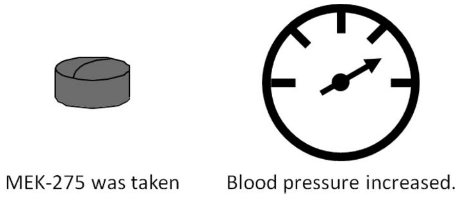


Figure 2. Sample stimuli used in Experiment 1. Medications were either taken or not, and some physiological outcome (e.g., blood pressure increases) either occurred or not.

Periodically, instead of progressing to the next trial, participants were asked to reconsider the current trial and select an interpretation of the observed events. The instructions were carefully written to indicate that participants were providing an interpretation of only a specific trial and not an interpretation of the overall causal strength up to that point. Specifically, learners were told, "you just observed a patient with the following information," and were asked to "choose the explanation that best describes what happened" from a list of choices. This only ever occurred on trials from Cell A (cause present, effect present) or B (cause present, effect absent). Because some models of learning (e.g., RW) make no substantive predictions about the influence of Cells C and D, these observations were not probed.

When asked to consider an observation from Cell A, participants were asked to choose between (a) "[medication] caused [symptom]," (b) "it is pure coincidence that [symptom] occurred after taking [medication]," or (c) "for some reason, the [medication] failed to prevent [symptom]," indicating a positive, neutral, or negative interpretation, respectively (see Table 1). For Cell B observations, participants were asked to choose between (a) "[medication] prevented [symptom]," (b) "it is pure coincidence that [symptom] did not occur after taking [medication]," or (c) "for some reason, the [medication] failed to cause [symptom]," indicating a negative, neutral, or positive interpretation, respectively (see Table 1).

In each condition, participants observed 64 trials consisting of 16 trials of each cell type as summarized in Figure 3. These trials were presented in two different orders: positive–negative or negative–positive. The positive–negative trial sequence presented the majority of positive evidence first (i.e., 14 of the 16 Cell A trials and 14 of the 16 Cell D trials), followed by the majority of negative evidence (i.e., 14 of the 16 Cell B trials and 14 of the 16 Cell C trials). The negative–positive trial sequence reversed the order of these two blocks. The trials within each sequence were presented in a quasi-randomized order to evenly distribute the different types of trials (see Figure 3). In particular, within each of the positive and negative blocks, the first two interpretation query trials were presented somewhere (random) after the fourth trial but before the 13th trial of the block, and the second two interpretation query trials were presented somewhere (random) after the 20th trial but before the 29th trial of the block. For each participant, two sets of stimuli were used for the two conditions. The order of the two conditions and the assignment of stimuli to the two conditions were counterbalanced across participants.

[5] Flexible interpretations per se can be accommodated by the power PC theory. (We thank Marc Buehner for pointing this out.) Yet the flexible influence of covariation information would pose a more direct challenge to the power PC theory.

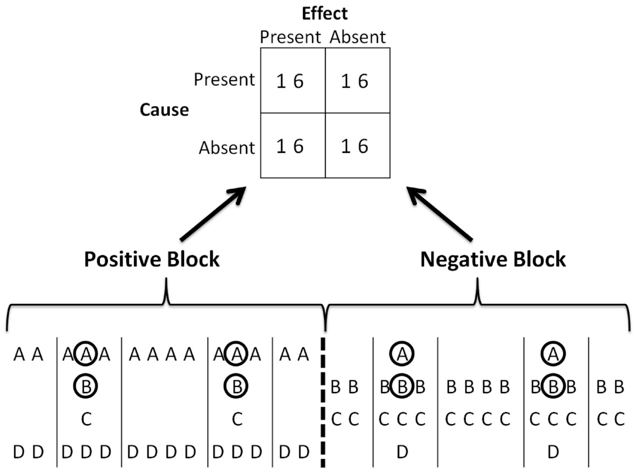


Figure 3. Trial sequences utilized in Experiment 1. The sequence was constructed as follows. Positive blocks began with two trials from Cell A and two from Cell D (randomly ordered). These were then followed by a series of eight trials: three each from Cells A and D and one each from Cells B and C (all eight randomly ordered). Within this eight-trial series, participants interpreted the B trial and one of the A trials, as indicated by circles in the figure. This was then followed by another series of eight trials: four each from Cells A and D (randomly ordered). These were then followed by another series of eight trials: three each from Cells A and D and one each from Cells B and C (all eight randomly ordered). Within this eight-trial series, participants again were asked to interpret the A trial and one of the B trials, as indicated by circles in the figure. The sequence then ended with two observations from Cell A and two from Cell D (randomly ordered). The top portion of the figure summarizes the overall contingency collapsed across the two blocks.

After viewing the entire set of trials, participants rated the causal strength of the medication by judging "the extent to which [medication] influences [symptom]" on a –100 ("[medication] prevented [symptom]") to 100 ("[medication] caused [symptom]") scale where 0 was labeled as "[medication] had no influence on [symptom]."

Results

Analyses of interpretations. As shown in Table 2, an average of 61.9% of interpretations differed from the way classic models of causal induction use these trials to modify causal beliefs (i.e., bold values in Table 2, such as negative or neutral interpretations of Cell A or joint presence of cause and effect). These interpretations occurred both for Cell A (M = 60.5%) and Cell B (M = 63.3%) and also across the positive–negative condition (M = 60.2%) and the negative–positive condition (M = 63.7%). Consistent with the expectation-based account, the inconsistent interpretations tended to occur more frequently in the second block (M = 68.0%) than in the first block (M = 55.9%), presumably because people would have a stronger expectation by the time they completed observing the first block and moved into the second block. Below, we offer comparative analyses to show interpretational flexibility.

To statistically evaluate participants' interpretations, we recoded each of the positive, negative, and neutral interpretation statements. Positive judgments (i.e., choosing Option a for Cell A and Option c for Cell B; see the Method section above) were scored as 1. Negative judgments (i.e., choosing Option c for Cell A and Option a for Cell B) were scored as −1. Neutral judgments appealing to pure coincidence were scored as 0. These scores, broken down by block and order, are shown in Figure 4. A 2 (order: positive–negative vs. negative–positive) × 2 (block: first half vs. second half) repeated measures analysis of variance (ANOVA) found no main effect of block but a significant main effect of order, $F(1, 31) = 57.30$, $p < .0001$, and a significant interaction effect, $F(1, 31) = 98.97$, $p < .0001$. Both of these significant effects are explored further.

We first considered judgments made in the first half of the trial sequences (i.e., during the positive half of the positive–negative order and the negative half of the negative–positive order). If our participants rigidly interpreted the covariation information, then these judgments should be equivalent because participants were always asked to interpret identical observations (i.e., two from Cell A and two from Cell B). To the contrary, they were significantly different from each other, $t(31) = 19.00$, $p < .0001$. Whereas first-half interpretations in the positive–negative order were significantly greater than zero ($M = .73$, $SD = .24$), $t(31) = 17.08$, $p < .0001$, those in the negative–positive order were significantly less than zero ($M = -.62$, $SD = .30$), $t(31) = 11.47$, $p < .0001$. That is, interpretations diverged in the first block such that they were generally consistent with the neighboring trials (a neighborhood effect henceforth).

However, second-half interpretations in the positive–negative ($M = -.007$, $SD = .50$) and negative–positive ($M = -.02$, $SD = .47$) orders did not differ from each other, $t(31) = 0.11$, *ns,* and did not differ from zero: positive–negative, $t(31) = 0.08$, *ns*; negative–positive, $t(31) = 0.28$, *ns.* This result suggests that the neighborhood effects were negated by expectations derived from the first half of the sequence. Thus, interpretations in a negative block were more negative when presented in the absence of prior experiences (i.e., the first half of the negative–positive order) than when preceded by a positive block (as in the second half of the positive–negative order), $t(31) = 5.51$, $p < .0001$. Similarly, interpretations

Table 2
Average Percentages of Choices for Each Interpretation in Experiment 1A

Condition and block	Generative	Neutral	Preventive
Positive–negative condition			
Positive block (first)			
Cell A	93.75	**6.25**	**0**
Cell B	**53.13**	**45.31**	1.56
Negative block (second)			
Cell A	28.13	**54.69**	**17.19**
Cell B	**23.44**	**40.63**	35.94
Negative–positive condition			
Negative block (first)			
Cell A	0	**57.81**	**42.19**
Cell B	**0**	**18.75**	81.25
Positive block (second)			
Cell A	35.94	**39.06**	**25.00**
Cell B	**12.50**	**59.38**	28.13

Note. Frequencies in bold represent interpretations that are inconsistent with the way classic causal learning models use covariation to modify causal beliefs.

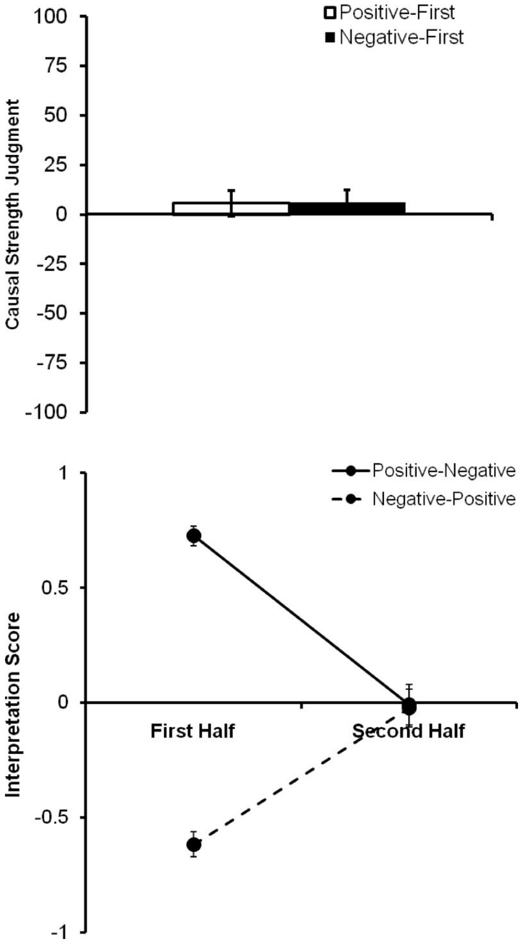


Figure 4. Experiment 1A interpretation results. Overall, participants exhibited neither recency nor primacy. Causal strength judgments did not differ depending on order. Interpretation judgments in the second half of the sequence show a similar pattern. Error bars represent ±1 standard error of the mean.

in a positive block were more positive when presented in the absence of prior experiences (i.e., the first half of the positive–negative order) than when preceded by a negative block (as in the second half of the negative–positive order), $t(31) = 7.77$, $p <$ .0001.

Relationship between interpretation judgments and causal strength judgments. Next, we examined whether our learners' interpretation judgments were linked to their learning as suggested by the expectation bias account. As shown in Figure 5, participants' judgments as a whole exhibited neither a primacy (e.g., Dennis & Ahn, 2001) nor a recency effect (e.g., López et al., 1998).[6] Judgments were near zero in both the positive–negative ($M = 5.56$, $SD = 35.99$) and the negative–positive ($M = 6.19$, $SD = 34.77$) conditions and did not differ from each other, $t(31) =$ .06, *ns.*

More importantly, we examined whether variance in individuals' causal strength judgments was related to variance in their interpretation judgments. Specifically, we evaluated whether the extent to which participants' interpretations were consistent with the first block correlated with the extent to which participants' causal strength judgments were consistent with the first block (i.e., the amount of primacy effect).

To do so, we computed (a) the extent to which the first-block interpretations were consistent with the first-block observations (i.e., measure of the neighborhood effect) and (b) the extent to which the second-block interpretations were consistent with the first-block observations (i.e., measure of the expectancy-based effect). Then, we examined the extent to which each of these measures correlated with (c) the magnitude of the primacy effects in participants' causal strength judgments. Appendix B explains how we computed these three measures, using concrete examples. Then, we performed a multiple regression analysis using (a) and (b) as predictors of (c). The results of this analysis indicate that learners' interpretation judgments were highly predictive of their causal strength judgments, $F(2, 29) = 6.56$, $p <$ .005. Inspection of the beta values reveals that it was only interpretations from the second block of the sequence that were related to causal strength judgments, $t(29) = 2.51$, $p <$ .05; first-block interpretations showed no such relationship, $t(29) = 1.53$, $p =$ .14. In other words, causal strength judgments were particularly related to how learners reacted to the second, conflicting block of observations in the order effect paradigm, rather than the neighborhood effects found in the first block. Those learners whose second-half interpretations were strongly consistent with the first block of the sequence also provided causal strength judgments that reflected the first block of the sequence. In contrast, learners whose second-half interpretations were consistent with the second block of the sequence tended to provide causal strength judgments that reflected the second block of the sequence.

To illustrate the relationship between interpretation judgments and causal strength judgments more vividly, we performed a median split of participants based on their causal strength score ($Mdn = 1.25$). This divides participants into those whose causal strength judgments generally exhibited primacy and those who generally exhibited recency. Figure 4 displays the interpretation judgments of these two groups. This graph illustrates how the two groups' interpretations differed. Those learners who exhibited more of a primacy effect provided second-half interpretation judgments that were more consistent with the first half of the trial sequence. In contrast, those learners who exhibited more of a recency effect provided second-half interpretations that were more consistent with the second half of the trial sequence.

Discussion

The current results suggest that people's causal learning may not treat covariation information statically, as traditionally implemented. Observations from Cell A, for example, were often interpreted as being consistent with a preventative causal belief. Thus, it appears that people's interpretations of covariation information were influenced by several factors other than the cells of the covariation matrix. First, we observed neighborhood effects, in which interpretations were modulated by the observations in the

[6] Any difference between the current results and those of Dennis and Ahn (2001) may be attributable to the repeated interruption, which has been shown to reduce the amount of primacy effect (Marsh & Ahn, 2006).

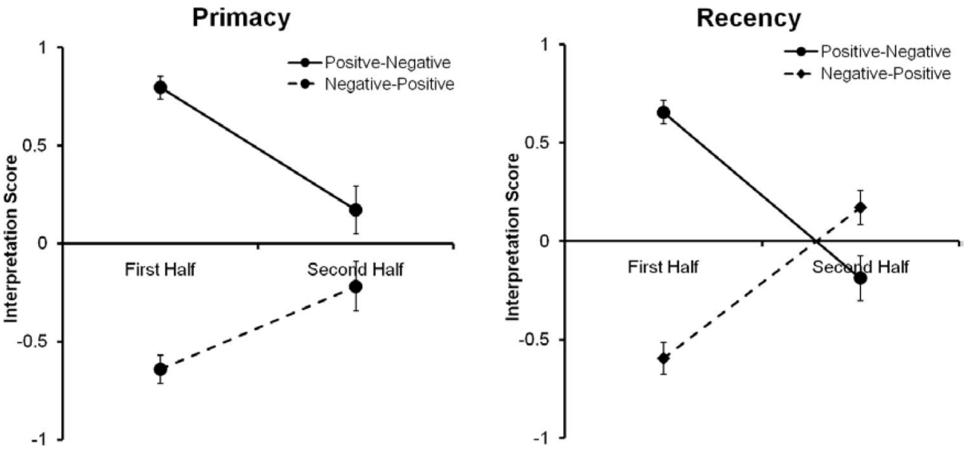


Figure 5. Interpretations after performing a median split on learners' causal strength judgments. Those in the primacy group exhibited greater primacy effect (more positive in the positive–negative condition and more negative in the negative–positive condition). Those in the recency group tended to exhibit the opposite pattern. Error bars represent ±1 standard error of the mean.

immediately surrounding sequence. Interpretations were more positive in positive blocks and more negative in negative blocks. Second, we observed more long-ranging effects of prior experience. The neighborhood effects observed early in learning were eliminated when learners had acquired conflicting prior experience.

One could argue that the participants were merely confused about the instructions and that, instead of providing an interpretation of a given trial, they were providing an overall estimate of causal strength up to that point in the sequence. For instance, according to this account, the second-half interpretations in the positive–negative order were made when the contingency was still somewhat positive (see Figure 3), which is why the second-half interpretations were somewhat positive. Disentangling such a possibility from the expectation-based account is difficult because we fully expect that individual interpretations are made according to learner's current beliefs about the strength of the causal relationship. We address this issue more directly later in the article (particularly Experiments 1C and 2).

For the next two experiments (Experiments 1B and 1C), we focused on the relationship between learners' interpretation judgments and their subsequent causal strength judgments. We found in Experiment 1A that our participants' interpretation judgments were not arbitrary but were connected to their learning in a principled manner. Our next two experiments manipulated either interpretation or causal strength judgments to see whether a similar relationship would continue to hold.

Experiment 1B

Experiment 1B manipulated the extent to which causal strength estimates are influenced by the presentation order of evidence. As discussed in the introduction, both recency (López et al., 1998) and primacy (Dennis & Ahn, 2001) effects are possible, and the recency effect is more likely when learners' working memory is overloaded (Marsh & Ahn, 2006). The current experiment used Marsh and Ahn's (2006) finding to elicit a recency effect. Participants proceeded through a learning task as in Experiment 1A, except that they were now required to perform a secondary task, designed to increase cognitive load, and thus induce a recency effect. As a result, we expected that the interpretation results would be similar to those of the participants who showed the recency effect in Experiment 1A (see Figure 4).

Method

Sixteen Yale University undergraduates participated for partial course credit. The stimuli and procedure were identical to Experiment 1A with one exception. While learners progressed through the trial sequence, they were required to count backward, by threes, from a large number (e.g., 286) provided by the experimenter. The counting task was performed during the entire learning sequence, including while making both interpretation and causal judgments. To ensure compliance, learners counted aloud and were told ahead of time that their counting would be monitored by experimenter stationed in an adjacent room (well within earshot).

Results and Discussion

When participants were required to perform a difficult secondary task, we observed a robust recency effect (see Figure 6). Causal strength judgments in the positive–negative order (M = −11.94, SD = 33.00) were much lower than the judgments in the negative–positive order (M = 19.25, SD = 29.68), $t(15) = 2.54$, $p < .05$. The critical question was whether this manipulation also influenced learners' interpretations.

Table 3 illustrates learners' interpretation judgments, again broken down by order and block. An average of 63.5% of interpretations were inconsistent with traditional role of covariation information (i.e., bold values in Table 3). Such interpretations occurred both for Cell A (M = 59.8%) and for Cell B (M = 67.2%), across the positive–negative condition (M = 63.8%) and the negative–positive condition (M = 63.3%), and in both the first block (M = 58.6%) and the second block (M = 68.5%).

For more fine-level statistical analyses, the three types of judgments were scored as before (positive = 1, negative = −1, neutral = 0), and the mean scores for each order, broken down by

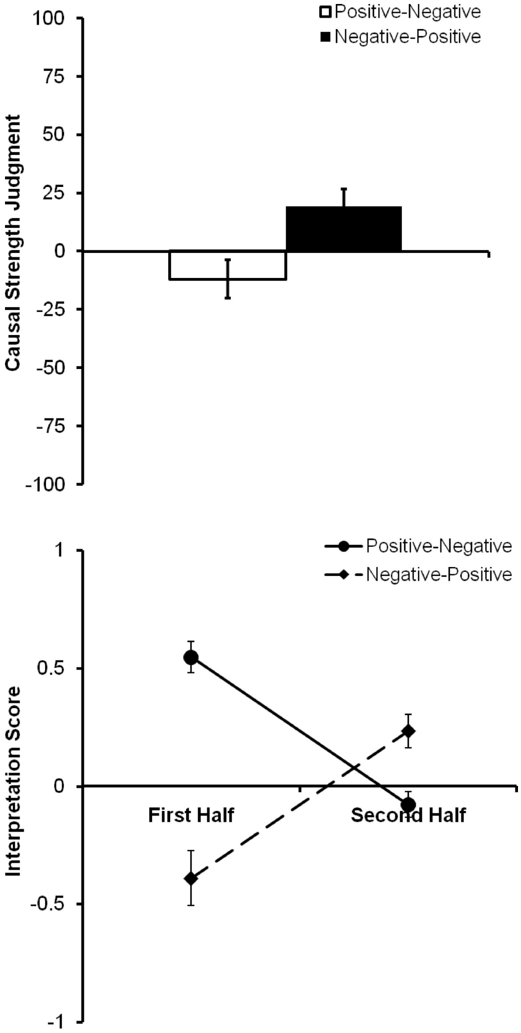


Figure 6. Experiment 1B results. Overall, participants exhibited recency effects. Causal strength judgments were more negative in the positive–negative condition and more positive in the negative–positive condition. Interpretation judgments in the second half of the sequence show a similar pattern. Error bars represent ±1 standard error of the mean.

block, are shown in Figure 6. A 2 (order: positive–negative vs. negative–positive) × 2 (block: first half vs. second half) repeated measures ANOVA found both a significant main effect of order, $F(1, 15) = 10.00$, $p < .01$ and an interaction between order and block, $F(1, 31) = 81.08$, $p < .0001$. As in Experiment 1A, interpretations during the first half of the sequence were consistent with the local context (a neighborhood effect) and thus differed across the two orders, $t(15) = 6.71$, $p < .0001$.

Where the current results diverge from Experiment 1A is in the second-half interpretations. Recall that, in Experiment 1A, second-half interpretations were near zero and identical across the two orders. In contrast, second-half interpretations in the positive–negative order ($M = -.08$, $SD = .22$) were significantly lower than those in the negative–positive order ($M = .23$, $SD = .28$), $t(15) = 3.18$, $p < .01$. That is, we observed neighborhood effects in the second half of the sequence; second-half interpretation judgments were consistent with the evidence contained in the second half of the trial sequence.

Thus, using a dual-task paradigm to induce a recency effect in learners' causal strength judgments, participants' interpretation judgments in the second block became less affected by the information presented in the first block. Learners in the positive–negative order made more negative interpretations in the second half and then provided primarily negative causal strength judgments. Learners in the negative–positive order made more positive interpretations in the second half and then provided primarily positive causal strength judgments.

Finally, we examined more directly the relationship between individual learners' interpretations and their causal strength judgments. As in Experiment 1A, we computed the extent to which interpretations and causal strength judgments were expectancy based, using the composite scores for each judgment (see Appendix B). These scores were then entered into a multiple regression model with participants' causal strength judgments as the dependent variable. Just as in Experiment 1A, causal strength judgments were strongly predicted by the extent to which the second-half interpretations were consistent with the first-half covariation, $t(13) = 2.57$, $p < .05$, but not by the extent to which the first-half interpretations were consistent with the first-half covariation, $t(13) < 1$. This result suggests that, despite modulating both causal strength and interpretation judgments, the secondary task did not eliminate the relationship between these judgments. That is, even with our experimental manipulation, we continued to observe a reliable relationship between individuals' interpretations, particularly those in the second half of the sequence, and causal strength judgments. This again suggests that there is a strong connection between the processes that underlie interpretation and learning.

Experiment 1C

Experiment 1C modulated learners' interpretations, instead of causal strengths. We introduced novel causes during specific, contradictory observations. For example, in the second half of a negative–positive sequence, observations from Cell A (i.e., cause

Table 3
Average Percentages of Choices for Each Interpretation in Experiment 1B

Condition and block	Generative	Neutral	Preventive
Positive–negative condition			
Positive block (first)			
Cell A	71.88	**28.13**	**0**
Cell B	**43.75**	**50.00**	6.25
Negative block (second)			
Cell A	13.33	**73.33**	**6.67**
Cell B	**16.67**	**36.67**	40.00
Negative–positive condition			
Negative block (first)			
Cell A	18.75	**46.88**	**34.38**
Cell B	**6.25**	**25.00**	68.75
Positive block (second)			
Cell A	50.00	**43.75**	**6.25**
Cell B	**12.50**	**78.13**	9.38

Note. Frequencies in bold represent interpretations that are inconsistent with the way classic causal learning models use covariation to modify causal beliefs.

present, effect present) were accompanied by a second, novel cause. Although the causal role of this second cause was left ambiguous, its addition was intended to bias the interpretation lent to these observations such that learners would be more likely to choose the negative, "for some reason, the [medication] failed to prevent [symptom]" interpretation. Unlike the preceding experiments, the novel cause should provide learners with a salient explanation or an excuse for the aberrant observation, essentially transforming the abstract "for some reason" into something rather concrete. We expected that this modification would lead to second-half interpretations even more consistent with the first half of the sequence. As a result, we also predicted that causal strength judgments would also exhibit greater primacy.

Method

Seventeen Yale University undergraduates participated for partial course credit. The stimuli and procedure were similar to Experiment 1A with a few exceptions. Participants were presented with information about two medications simultaneously. Participants were told that their task was to learn about only one of the two causes, the target cause, and that they would not have to judge the other cause, the alternative cause. The target cause behaved as in Experiment 1A (see Figure 3). The alternative cause was only present in the second half of the trial sequence and then only during observations that were inconsistent with the first block and on which the target cause was present. Thus, the alternative cause was present during all observations from Cell B in the negative half of the positive–negative order and during all observations from Cell A in the positive half of the negative–positive order.

The alternative cause was introduced in the second half, on inconsistent trials, to modulate learners' interpretations so as to reflect the first half of the trial sequence more than the second half. For instance, after observing the first half of the positive–negative order and developing the expectation that there is a positive causal relationship, observing the target cause and the alternative cause in the absence of the effect (i.e., Cell B) is likely to elicit the interpretation that the alternative cause is to be blamed for the failure to produce the effect. We predicted that our experimental manipulations of learners' interpretation judgments would also elicit more expectation-based causal strength judgments.

Results and Discussion

Table 4 shows the mean percentage of choices for interpretation judgments. As before, a bulk of trials (*M* = 62.13%) were not interpreted as traditionally implemented. Such interpretations frequently occurred for both Cell A (*M* = 58.8%) and Cell B (*M* = 65.4%), across the positive–negative order (M = 55.1%) and the negative–positive order (*M* = 69.1%), and for both the first block (*M* = 60.3%) and the second block (*M* = 64.0%).

Figure 7 shows participants' mean interpretation scores (1 = positive, 0 = neutral, −1 = negative). A 2 (order: positive–negative vs. negative–positive) × 2 (block: first half vs. second half) repeated measures ANOVA revealed a significant main effect of order, $F(1, 15) = 29.01$, $p < .0001$, but no interaction between order and block, $F(1, 32) < 1$, *ns.* First-half interpretations were again consistent with the surrounding context. Those in the positive–negative order were significantly greater than zero (*M* = .59, *SD* = .31), $t(16) = 7.94$, $p < .0001$, whereas those in the negative–positive order were significantly less than zero (*M* = −.49, *SD* = .42), $t(16) = 4.78$, $p < .0001$. In contrast, second-half interpretations were overwhelmingly consistent with the first half of the trial sequence (and thus inconsistent with the surrounding context). Second-half judgments from the positive–negative order were significantly greater than zero (*M* = .57, *SD* = .38), $t(16) = 6.18$, $p < .0001$, whereas judgments from the second half of the negative–positive order were significantly less than zero (*M* = −.50, *SD* = .47), $t(16) = 4.41$, $p < .0005$. Averaging across blocks, the overall difference between the two orders was also significant, $t(16) = 8.46$, $p < .0001$, suggesting that our manipulation had the desired effect on learners' interpretations. The critical question was whether this manipulation would also influence learning as measured by participants' causal strength judgments.

Table 4
Average Percentages of Choices for Each Interpretation in Experiment 1C

Condition and block	Generative	Neutral	Preventive
Positive–negative condition			
Positive block (first)			
Cell A	91.18	**5.88**	**2.94**
Cell B	**38.24**	**52.94**	8.82
Negative block (second)			
Cell A	67.65	**32.35**	**0**
Cell B	**58.82**	**29.41**	11.76
Negative–positive condition			
Negative block (first)			
Cell A	0	**52.94**	**47.06**
Cell B	**8.82**	**32.35**	61.76
Positive block (second)			
Cell A	5.88	**41.18**	**52.94**
Cell B	**5.88**	**35.29**	58.82

Note. Frequencies in bold represent interpretations that are inconsistent with the way classic causal learning models use covariation to modify causal beliefs.

As can be seen in Figure 7, participants' causal strength judgments exhibited a strong primacy effect. Causal strength judgments in the positive–negative order were significantly greater than zero (*M* = 45.41, *SD* = 28.11), $t(16) = 6.66$, $p < .0001$, whereas judgments in the negative–positive order were significantly less than zero (*M* = −31.35, *SD* = 35.98), $t(16) = 3.59$, $p < .005$. The difference between two orders was significant, $t(16) = 8.06$, $p < .0001$.

Finally, using multiple regression over the composite scores for interpretation judgments and causal strength judgments as explained in Experiment 1A and Appendix B, we again tested whether individuals' interpretations would predict their causal strength judgments. Just as in the previous two studies, causal strength judgments were predicted by the extent to which second-half interpretations were consistent with the first block, $t(14) = 2.32$, $p < .05$, but not by the extent to which the first-half interpretations were consistent with the first block, $t(14) < 1$.

To summarize, by providing an excuse, participants could more readily explain away aberrant trials in the second block, and we found much stronger expectancy-based interpretations in the second block. At the same time, participants' causal strength judgments reflected the first half of the sequence even more than the second half.

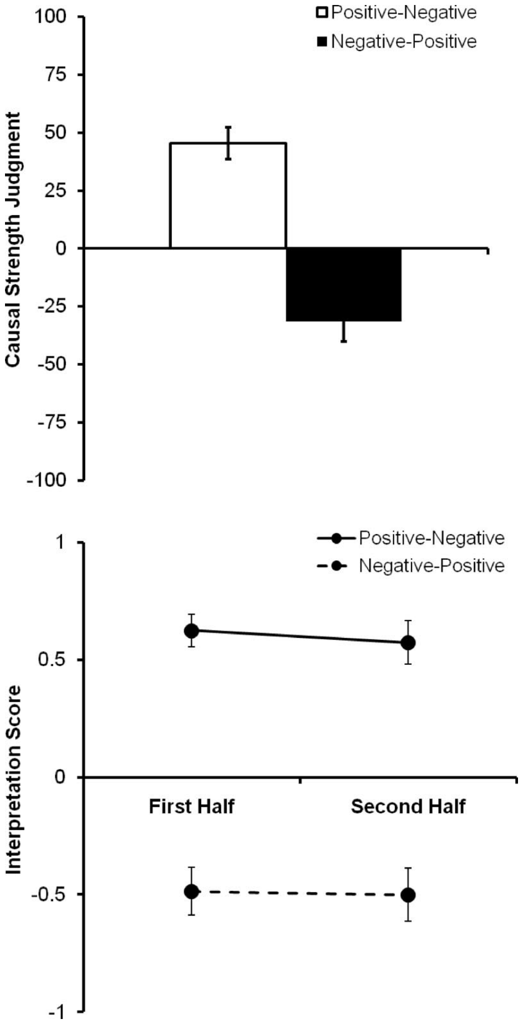


Figure 7. Experiment 1C results. Causal strength judgments were more positive in the positive–negative condition and more negative in the negative–positive condition. Interpretation judgments in the second half of the sequence show a similar pattern. Error bars represent ±1 standard error of the mean.

Experiment 1C also undermines one possible alternative account to our claim that interpretations derive causal strength judgments. This counterargument claims that the interpretation judgments merely reflect Δ*P* experienced so far, or causal power (Cheng, 1997) estimated up to that point. For instance, by the time participants provided the last interpretation judgments in the first block, the overall Δ*P* and causal power were about 0.71 and 0.83, respectively, in the positive–negative sequence and −0.71 and −0.83, respectively, in the negative–positive sequence. By the time they provided the last interpretation judgments in the second block, the overall ΔP and causal power were about 0.07 and 0.13, respectively, in the positive–negative sequence and about −0.07 and −0.13, respectively, in the negative–positive sequence. Thus, the difference between the two orders became smaller by the time participants experienced the second block, which mirrors the interpretation judgments observed in Experiments 1A and 1B. In Experiment 1C, we used identical trials (i.e., identical Δ*P* and causal power), but the interpretation results observed in Experiment 1C do not at all reflect this reduced difference between the two conditions in the second block. Although the covariation between the target cause and the effect dramatically changed in the second half of the sequence, just as in Experiments 1A and 1B, the interpretation judgments from the first and the second halves were indistinguishable.

Experiment 2

Experiment 1 demonstrated that observations from Cell A can be interpreted as evidence for an inhibitory causal relationship and that observations from Cell B can be interpreted as evidence for a generative causal relationship. Experiment 2 tested a stronger version of our proposal; Cell A can act to decrease a causal strength and Cell B can act to increase a causal strength. That is, covariation information can exert influences on overall causal strength judgments in ways that oppose traditionally implemented roles.

The design is illustrated in Figure 8. In the control conditions, the sequence consisted of two blocks (one positive and one negative), but the second block contained 10 fewer trials than in Experiment 1 (see the trials without brackets in Figure 8). More specifically, the control condition of the positive–negative sequence contained five fewer B and five fewer C trials in the second half, making the overall contingency positive ($\Delta P = 0.2$). Similarly, the control condition of the negative–positive sequence contained five fewer A and five fewer D trials in the second half, making the overall contingency negative ($\Delta P = -0.2$). In the experimental conditions, the positive–negative and negative–positive sequences were identical to Experiment 1 (except for alternative causes; see below). Thus, the overall contingency of each sequence was zero.

Importantly, in the experimental conditions, on these extra observations (i.e., the bracketed trials in Figure 8), a second, alternative cause was present. As in Experiment 1C, these were inserted to elicit more nontraditional interpretations of covariation information by reminding learners that there may be alternative causes that can be blamed for aberrant observations.

According to many traditional models of causal learning, including RW, Δ*P*, and power PC, judgments after the positive–negative sequence should be more negative in the experimental condition than in the control condition, and judgments after the negative–positive sequence should be more positive in the experimental condition than in the control condition. For Δ*P* and power PC, this is because Δ*P* is 0 for both sequences in the experimental condition, whereas, in the control condition, Δ*P* is 0.2 and −0.2 of the positive–negative and negative–positive sequences, respectively (see Figure 8 for the summary). One may argue that presence of alternative causes may create a distinctive context and that a learner might conditionalize on the presence or absence of alternative causes (e.g., Spellman, 1996). If participants treat trials with alternative causes separately or ignore them, Δ*P* and power PC predict that the control and experimental conditions are identical. However, these models can never predict that the experimental condition of the positive–negative sequence, which contains more B and C trials than its control condition, would result in higher causal judgments than its control condition. Likewise, they can never predict that the experimental condition of the negative–positive sequence, which contains more A and D trials than its

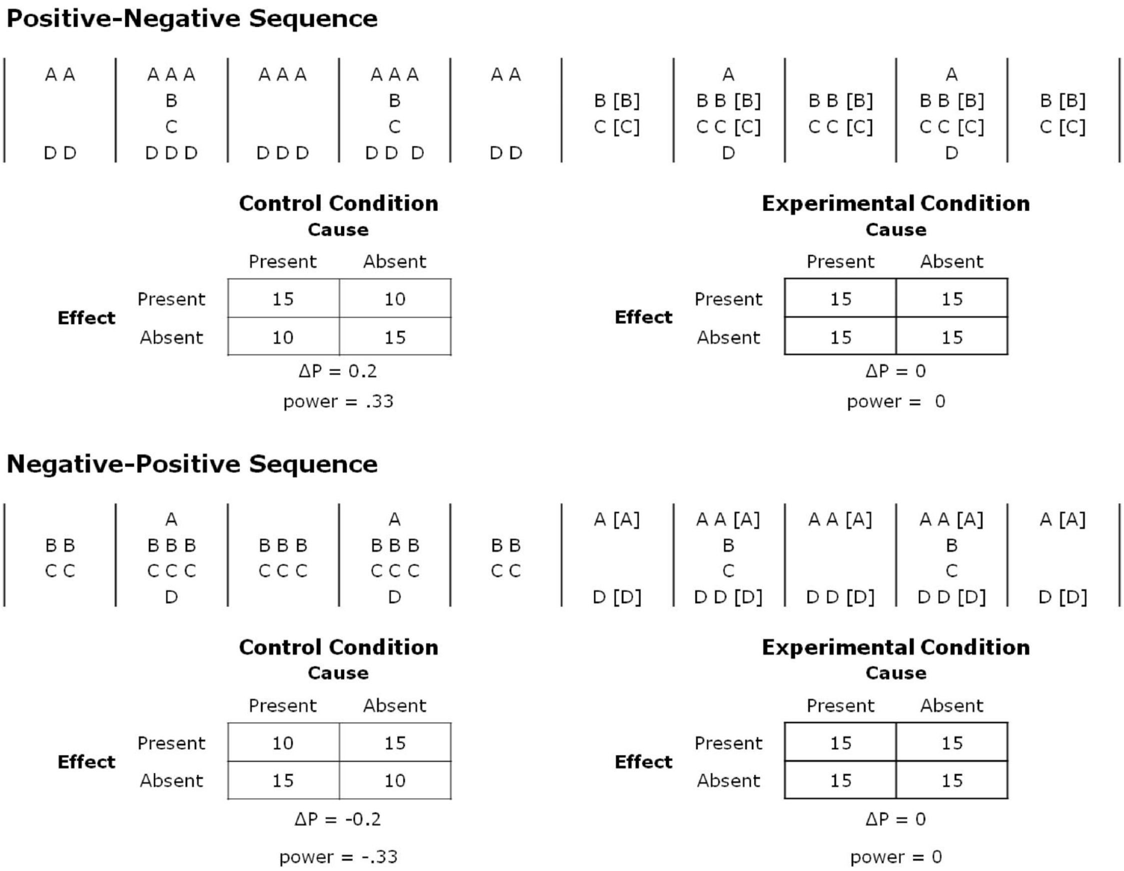


Figure 8. The design of Experiment 2. The control condition included sequences with slightly longer first halves so that the overall covariation (ΔP) would be more reflective of the evidence presented at the beginning of the sequence than the evidence presented at the end. The experimental condition added additional observations to the second half of the sequences (placed in brackets). These additional observations also included a second, alternative cause (see text for further details). Importantly, these extra observations changed the overall covariation of both sequences to be zero.

control condition, would result in lower causal judgments than its control condition.

RW makes similar predictions. The experimental condition contains more B and C trials in the second half than the control condition, and therefore, regardless of the parameter values, the final associative strength of the experimental condition can never be higher than that from the control condition of the positive–negative sequence. Similarly, RW would never predict the control condition of the negative–positive sequence to be lower than that from its experimental condition.

In contrast, we predicted that the exact opposite would occur. Suppose a learner is progressing through a positive–negative sequence. The first half of the sequence creates a belief that the cause is strongly generative in nature. When encountering the second, negative half of the sequence, learners in the control condition will process this negative information in light of their existing hypothesis, blunting this information's influence. Learners in the experimental condition will be confronted with even more negative evidence. However, some of this evidence will occur in the presence of an alternative cause, allowing learners to interpret these observations as being consistent with their existing, positive hypothesis (e.g., Experiment 1C). We further expected that the presence of an alternative cause on these few trials would encourage similar interpretations of B and C trials without alternative causes throughout the remainder of the sequence (or perhaps retrospectively), ultimately leading to more positive beliefs. The opposite would occur with the negative–positive sequence. Thus, adding observations that, according to the traditional accounts, contradict the initial observations may be able to paradoxically reinforce learners' initial hypothesis.

Method

Participants. Forty-four Stony Brook University (Stony Brook, NY) undergraduates participated for partial course credit.

Materials and procedure. The stimuli and procedure were similar to those used in Experiment 1, although participants were no longer asked to provide interpretations. Stimuli consisted of novel medications (e.g., DJE-143), each of which could have some causal influence on granulocytes (described as "a substance produced in people's blood"). Participants were told that it was their job to determine what influence the medications had on granulocytes. Each participant was again exposed to two different trial sequences, each of which instantiated two different orders: positive–negative and negative–positive. Figure 8 shows the actual frequencies and sequences. The trials within each sequence were presented in a quasi-randomized order to evenly distribute the different types of trials.

The critical manipulation in Experiment 2 occurred during the second half of each sequence. Participants in the control condition received sequences consisting of trials without brackets in Figure 8. In contrast, participants in the experimental condition received an additional 10 observations in the second half of the sequence (i.e., bracketed trials in Figure 4). The experimental condition of the positive–negative sequence included five additional Cell B trials and five additional Cell C trials presented in the second half of the sequence, and the experimental condition of the negative–positive sequence included five additional Cell A trials and five additional Cell D trials presented in the second half of the sequence. (These additional trials made the sequences of the experimental conditions identical to those in Experiment 1.) On these extra trials, a second, alternative cause was also always present. This alternative cause was not mentioned on any other trials during the sequence (i.e., not explicitly present, not explicitly absent). To make sure that all learners (in both conditions) had equivalent expectations about alternative causes for the trials that were shared between the two conditions, the initial task instructions specified, “If we know whether a given patient is taking other medication, that information will be presented to you. If there is no information about other medication, the patient may or may not be taking other medication, we simply don’t know.”

After viewing the entire set of trials, all participants rated the causal strength of the medication. Specifically, participants were asked to, “judge the extent to which [medication] influences granulocytes.” Responses could range from –100 (“[medication] prevented granulocytes”) to 100 (“[medication] caused granulocytes”), with 0 labeled as “[medication] had no influence on granulocytes.”

The manipulation of the experimental and control conditions was a between-subject variable ($N = 22$ in each condition) and the order (positive–negative vs. negative–positive sequence) was a within-subject variable. For each participant, a different medication–symptom pair was utilized in the two different order sequences. The two orders and the assignment of stimuli were counterbalanced across participants.

Results and Discussion

Figure 9 shows the results. A 2 (experimental vs. control) × 2 (order) ANOVA with repeated measure on the latter factor found a significant main effect of order, $F(1, 82) = 13.66$, $p < .0005$, and, more critically, a significant interaction effect, $F(1, 82) = 12.99$, $p < .001$. As predicted, this interaction was obtained because the positive–negative sequence elicited more positive judgments in the experimental condition ($M = 55.73$, $SD = 28.94$) than in the control condition ($M = 25.23$, $SD = 39.05$), $t(42) = 2.93$, $p < .01$, whereas the negative–positive sequence elicited more negative judgments in the experimental condition ($M = -43.91$, $SD = 34.02$) than in the control condition ($M = -24.36$, $SD = 28.84$), $t(42) = 2.05$, $p < .05$.

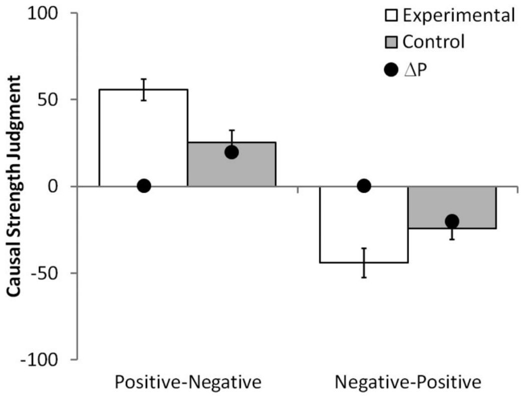


Figure 9. Experiment 2 results. The experimental condition added observations to the second half of each sequence (additional observations from Cells B and C in the positive–negative sequence and additional observations from Cells A and D in the negative–positive sequence). The solid circles represent the predictions of the ΔP model (causal power predictions are similar). As can be seen, the pattern exhibited by ΔP indicates that the addition of Cell B and C observations will lower causal strength judgments and that the addition of Cell A and D observations will raise causal strength judgments. Participants exhibited the opposite pattern. Error bars represent ±1 standard error of the mean.

This pattern conforms to our predictions and provides strong evidence that the four covariation evidence types do not always exert the traditional influence on learning. Adding observations on which the target cause was present and the effect was absent (Cell B) and observations on which the target cause was absent and the effect was present (Cell C) resulted in more generative causal strength judgments. Conversely, adding observations on which the target cause and effect were both present (Cell A) and observations on which the target cause and effect were both absent (Cell D) resulted in more preventative causal strength judgments.[7]

[7] One may argue that the additional trials marked with potential alternative cues could have implied a change in context. One possibility is that participants could have believed that only the trials with potential alternative cues had different contexts. As discussed at the beginning of Experiment 2, however, if only these trials were ignored by the participants, the experimental and the control conditions would be equivalent, so this possibility cannot account for the current results. The second possibility is that the occasional presence of a potential alternative cause could have resulted in the wholesale ignorance of the second half of the sequences. In this case, the causal powers according to power PC in the first half of the sequences would be 0.86 in the positive–negative sequence and −0.86 in the negative–positive sequence. The results from both Experiment 1C (0.45 and −0.31, respectively) and the experimental condition of Experiment 2 (0.55 and −0.43, respectively) deviate substantially from these predictions. Furthermore, the judgments deviate in the direction (toward zero) that would be expected if learners did in fact processes the second halves of these sequences. Last, we note that previous authors (e.g., Glautier, 2008) have suggested that a change in context might explain recency effects in causal learning. That is, these authors proposed that a change in context halfway through a learning sequence would cause learners to ignore the first half (rather than the second half, as suggested by a reviewer). Given the amount of disagreement between these various accounts, future work will be needed to more thoroughly explore the relationship between them.

General Discussion

We have suggested that an integral part of causal learning is the subjective processing of individual pieces of covariation information. The current study sought to provide support for this proposal in two ways. In Experiment 1, we had learners provide explicit interpretations during the course of learning. These results illustrate several important points. First, interpretations of identical observations varied depending on the immediately surrounding context. For example, interpretations made in the midst of a strongly positive block were more positive than interpretations of identical observations made in the midst of a strongly negative block. Second, interpretations varied depending on the expectations learners developed at the beginning of learning. For instance, learners interpreted observations more negatively if learners had previously encountered a strongly negative sequence of observations than if they had not. Last, we found that these interpretations were systematically related to causal strength judgments. Individual learners' interpretations idiosyncratically predicted their own strength judgments. In addition, we found interpretations and causal strength judgments to be linked at the group level using experimental manipulations designed to modulate causal strength judgments (Experiment 1B) and interpretations (Experiment 1C).

In Experiment 2, we went even further and demonstrated that the addition of supposedly positive covariation information could result in more negative causal strength judgments and vice versa. Compared to the control conditions, the experimental conditions had a small number of extra trials that were inconsistent with the first half of the sequence and included a second, alternative cause. The presence of the alternative cause allowed learners to interpret these observations in a way that was consistent with their existing hypotheses (e.g., Experiment 1C). Presumably, because of these unconventional interpretations, causal strength judgments were altered in ways that did not correspond to the assumptions of traditional learning theories. This experiment is the first empirical demonstration that additional presentation of Cells B and C can result in more generative causal strength judgments and that additional presentation of Cells A and D can result in more preventative causal strength judgments.

The current study helps to explain previous reports of primacy effects in causal learning (Chapman & Chapman, 1969; Dennis & Ahn, 2001; Yates & Curley, 1986). Our results demonstrate that the hypotheses learners carry with them influence their interpretations, which, in turn, bias the changes that learners make to their hypotheses. Learners who have just observed a set of evidence suggesting a strong, generative causal relationship process subsequent evidence differently than learners who have different prior experience. This processing, under many conditions, operates such that evidence is processed to be consistent with the learner's current expectations. Such expectation-based processing acts to attenuate the influence of subsequent, conflicting evidence and tends to yield primacy effects.

Our results further explain the dependence of primacy effects on working memory (Marsh & Ahn, 2006), since working memory is required for expectations to fully influence the interpretation process. Without access to working memory, interpretations no longer bias data to be consistent with previously established expectations, and subsequent causal strength judgments reflect the relatively unbiased, recent data. Thus, despite the variety of order effects we observed in the current experiments (primacy, recency, and no order effect at all), our data suggest that they may all be explained with a single process.

Token Versus Type Causal Cognition

One aspect of causal cognition that has received surprisingly little attention is the relationship between reasoning about causal tokens and causal types. Causal types are those causal beliefs that describe a generality: Coffee causes alertness. This is the sort of causal belief that is typically explored in causal learning paradigms (Shanks et al., 1996). Causal tokens are those causal beliefs that describe a specific causal interaction taking place in a specific time and place: This morning's coffee made me alert. This is the sort of belief that is typically labeled as causal reasoning (Goldvarg & Johnson-Laird, 2001; Wolff, 2007; Wolff & Song, 2003) and is often framed as a problem to be solved (e.g., why are you alert this morning?). Despite having two separate literatures, there is no particular reason to think that reasoning about types and reasoning about tokens are unrelated.

The causal strength judgments our learners made after observing the entire sequence of observations were designed to elicit beliefs about the causal type. That is, these judgments asked about the causal pattern that held in general. The interpretations that learners make about individual trials are beliefs about the causal token. For example, the interpretation judgments in Experiment 1 asked for a causal description of a single episode. The neighborhood effects and the effect of order on interpretation judgments found in Experiment 1 suggest that learners' beliefs about causal types influenced their interpretation of the causal interactions on individual trials. Consistent with this, we observed that interpretation judgments, particularly those in the second half of the sequence, predicted subsequent causal strength judgments across individuals, possibly suggesting a mutual influence of type and token processing.

Last and perhaps most interesting, Experiment 2 strongly suggests that the interpretation of individual observations can influence causal strength judgments. In this experiment, we provided learners with a select number of trials in which an alternative cause was present. These trials were specifically designed to bias the causal explanation given to individual observations; learners were given an opportunity to excuse evidence that, according to the traditional interpretation of covariation information, contradicted the first half of the learning sequence. Our results demonstrate that, when given this opportunity, learners used the alternative cause to interpret the seemingly contradictory observations in atypical ways. As a result, these trials did not influence learners' subsequent causal strength judgments in the manner that has been traditionally assumed.

Given these results, we argue that the processing of causal types and the processing of causal tokens exert mutual influence on each other. This suggests that the processes that underlie what is generally labeled as causal reasoning and those that underlie causal learning may be heavily intertwined. This is a relationship that is currently absent from nearly all current causal learning models. One exception is our recent model, BUCKLE (bidirectional unobserved cause learning; Luhmann & Ahn, 2007), which describes how people learn about causes that they cannot observe. According to BUCKLE, learners engage in causal reasoning on each trial in an attempt to overcome missing information. Learning then operates by using the results of the causal reasoning as a replacement for the missing information. Our results suggest that something similar was going on in the current study. Instead of trying to figure out whether causes were present or absent, learners in the current study appear to have been engaged in generating explanations for individual observations. As in BUCKLE, conclusions drawn about individual observations apparently influenced subsequent learning. Of course, we do not currently have any theories about how learning might use explanations or interpretations as input because it has been assumed that unprocessed covariation is sufficient to capture learning behavior. Further work on this topic has the ability to both enrich and unify an understanding of causal learning, causal reasoning, and the relationship between these abilities.

Bayesian Inference

As discussed earlier, the two sets of classical models of learning cannot readily account for all aspects of our current results. What sort of process might explain both the interpretations and causal strength judgments we have reported here? A Bayesian approach seems immediately attractive. The hallmark of Bayesian inference is the ability to integrate newly observed data with existing beliefs or hypotheses. Here, we briefly review several of the most recent Bayesian models of causal learning, discuss their important theoretical contribution, and explain how they relate to our current findings.

Griffiths and Tenenbaum (2005) suggested that learners do not attempt to determine the strength of individual causal relationships but rather attempt to determine whether there exists a causal relationship (of any strength) between the potential cause and the potential effect. For example, when asked to evaluate the degree to which carbon dioxide causes global warming, Griffiths and Tenenbaum suggested that people's judgments can be thought to reflect a decision between a model of the world in which these two variables are causally connected and a model of the world in which the two variables are unrelated. To test this proposal, Griffiths and Tenenbaum formalized their proposal with a Bayesian model called causal support and fit the model to several data sets. Each set of data was collected by presenting a homogeneous set of positive covariation information to participants and eliciting causal strength judgments. The results demonstrated that, across data sets, participants' judgments tended to exhibit variability even when other models (e.g., power PC and ΔP) would suggest they should be constant. In general, the causal support model was able to account for many of these apparent deviations and provided better overall fits to the data.

Lu, Yuille, Liljeholm, Cheng, and Holyoak (2008) proposed a Bayesian extension of Cheng's power PC theory (Cheng, 1997). They proposed that people expect causes to be both necessary and sufficient to bring about their effects. That is, it is suggested that people have an aversion to causes that are not systematically followed by their effects (sufficiency) and an aversion to effects that are not systematically preceded by their causes (necessity). Lu, Yuille, et al. formalized this expectation as a set of Bayesian priors and compared the resulting model with an identical model having no preference for necessity or sufficiency (which is essentially just causal power as described by Cheng, 1997). The two models were then fit to two types of data. The first type of data was collected by presenting participants with homogeneous covariation information and then eliciting causal strength judgments. When the two models were fit to the causal strength judgments, there was little difference between the model that preferred necessity and sufficiency and the model that had no preference. The second type of data was collected by presenting participants with homogeneous covariation information and eliciting judgments of causal structure (cf. Griffiths & Tenenbaum, 2005). Here, the model that expected necessity and sufficiency provided significantly better fits to a variety of causal structure judgments.

Both Griffiths and Tenenbaum (2005) and Lu, Yuille, et al. (2008) put forth formal, Bayesian accounts of how covariation information is used to produce causal judgments. However, the causal support model was explicitly designed to model judgments of causal structure (rather than causal strength, which was elicited in the current study), and the prior expectation about necessity and sufficiency proposed by Lu, Yuille, et al. was only required to account for judgments of causal structure. Because of the focus on causal structure rather than causal strength and because both of these models focus on aggregate covariation data rather than individual pieces of covariation information, these models are not immediately applicable to the methods utilized in the current study or the order effects found in the current study.

Recently, there have been two models proposed (Lu, Rojas, et al., 2008; Lucas & Griffiths, 2010) that are much more related, at least in spirit, to the current study. Each of these models suggests that participants simultaneously learn two related things from covariation information. Learners must first determine how multiple causes combine their influence to bring about their effects. For example, causes may combine their effects additively as is the case when one takes two aspirin; two aspirin are twice as effective as one aspirin. Alternatively, causes may interact, as is the case when one mixes bleach and ammonia; despite neither being particularly dangerous on its own, the combination is highly toxic. On the basis of their beliefs about how causes combine their influence, learners must also determine the strengths of the various causes.

To evaluate these proposals, the authors applied them to experiments in which learners received two sets of covariation information. In the first, training phase, learners received data that provided hints as to how causes combined their influence to produce their effects. For example, one experimental condition included training data that, like the bleach and ammonia example, described a pair of causes that each failed to produce the effect on its own but that were able to produce the effect when paired together. Another condition included training data that, like the aspirin example, described a pair of causes that each produced the effect on its own and also produced the effect when paired together. To evaluate the influence of the training, all learners were then provided with an identical set of covariation information and asked to make causal strength judgments. Despite making causal strength judgments about identical covariation information, the different training conditions led to different causal judgments. For example, if one learned that causes do not interact (like aspirin), then observing a single cause that is not followed by its effect (i.e., Cell B) would suggest that that cause was not particularly strong. If one had instead learned that causes do interact (like bleach and ammonia), then seeing the same observation (e.g., ammonia not being dangerous) would leave its interactive causal strength ambiguous.

These more recent proposals (Lu, Rojas, et al., 2008; Lucas & Griffiths, 2010) describe a more sophisticated formal account (so-called hierarchical Bayesian inference) about how people extract causal beliefs from covariation information. Just as has been demonstrated in the current study, these models exhibit significant flexibility in how covariation information leads to causal beliefs. Specifically, these models are able to reach different causal conclusions based on identical covariation information. The flexibility exhibited by these models is directly tied to their assumption that people simultaneously learn how causes interact to produce their effects and learn causal strengths. Because a critical component of these models involves learning about how multiple causes combine their influence, however, they cannot be directly applied to the current study (which only ever had participants learning about a single cause). Yet these models clearly illustrate how the hierarchical Bayesian approach to causal learning can lead to a significantly greater flexibility than traditional models.

So far, we have reviewed a recent crop of Bayesian models that have each described substantially more sophisticated processes by which covariation information is transformed into causal beliefs. Although each of these models relies on prior beliefs to dictate exactly how causal learning proceeds, the specific models that have been proposed are not immediately applicable to our results for a variety of reasons. Below, we provide more general discussion of the reasons for the limitations.

First, as explained in the earlier results, much of the flexibility we observed in the current study seems to have been a consequence of a strong relationship between beliefs about types (e.g., does this cause produce the effect in general?) and beliefs about tokens (e.g., did this cause produce the effect on this occasion?). The flexibility exhibited by the newer Bayesian models is a consequence of beliefs about causal strength (type) and beliefs at a more abstract level. Griffiths and Tenenbaum's (2005) causal support model relies on beliefs about potential causal structure, Lu, Yuille, et al. (2008) modeled prior beliefs about the necessity and sufficiency of causes, and both Lucas and Griffiths (2010) and Lu, Rojas, et al. (2008) modeled beliefs about how causes interact. To account for the results of Experiment 1, focus would instead have to be directed to beliefs at a more specific (i.e., token) level.

Second, our results demonstrate the influence of a variety of different kinds of prior beliefs, and it is not yet clear how the entire set could be implemented within the Bayesian framework. Learners' interpretations were influenced by the proximal context, what we refer to as neighborhood effects. Observations surrounded by predominately negative covariation information tended to be judged negatively, whereas observations surrounded by predominately positive covariation information tended to be judged positively. In addition, learners' interpretations were influenced by the distal context. Observations preceded by a large block of negative covariation information tended to be judged negatively, whereas observations preceded by a large block of positive covariation information tended to be judged positively. These factors embody exactly the sort of influences used by Bayesian inferences but suggest that a process account will be necessary to fully account for our results. There have been Bayesian accounts of causal learning that attempt to account for the trial-by-trial dynamics of learning (e.g., Danks, Griffiths, & Tenenbaum, 2003; Lu, Yuille, et al., 2008), but many of these are extremely simplistic, typically applying standard inference iteratively after each trial. Only the proposal of Lu, Rojas, et al. (2008) exhibits the degree of sophistication likely necessary to deal with the current data.

The third factor is much more psychological in nature. In Experiment 1B, participants placed under cognitive load continued to be influenced by the local context but were not particularly sensitive to the more long-ranging influence of the preceding block of trials. To integrate this factor into a Bayesian account, one could presumably posit a resource-dependent process that is responsible for combining prior beliefs with current data. For example, simply holding prior beliefs in memory may be cognitively taxing and thus be compromised in individuals or situations where working memory is compromised.

Impression Formation

Our theorizing about the order effect paradigm has been substantially informed by prior work on impression formation. In particular, the order effect paradigm itself was taken from work originally done on impression formation (Asch, 1946). The paradigm is completely analogous to the current Experiment 1; the presentation order of identical sequences of information is manipulated without changing the information itself. Much like the current results, Asch's (1946) findings demonstrated the influence of prior expectations. If one's first experience with a stranger is predominantly positive, one will tend to overlook subsequent negative experience and judge the individual to be relatively positive, whereas the reverse presentation order will lead to relatively negative judgments (Asch, 1946). The similarity between the primacy effects reported in impression formation and causal learning is striking. Interestingly, there have also been reports of recency effects in impression formation tasks using an order effect paradigm. One factor that appears to partly determine whether one observes primacy or recency is the frequency with which participants are required to make judgments of the person under scrutiny (Stewart, 1965). Studies requiring more frequent judgments tend to report recency, whereas those that elicit judgments only at the end of the sequence tend to report primacy. Interestingly, this exact same pattern of data has been reported within the causal learning literature (Catena, Maldonado, Megias, & Frese, 2002; Collins & Shanks, 2002; Matute et al., 2002; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996).

Last, there is also evidence that impression formation is subject to the effects of the local context just as with the neighborhood effects reported here. In his landmark warm–cold study, Asch (1946) found that impressions could be strongly swayed by simply inserting either *warm* or *cold* into an otherwise identical list of characteristics. Those who were described with the *warm* list were judged as generally positive, whereas those described by the *cold* list were judged as generally negative. Asch suggested that this manipulation "did not simply add a new quality, but to some extent transformed the other characteristics" (Asch, 1946, p. 264). A subsequent experiment confirmed this by having participants provide and explain similarity judgments. For example, when someone was described as quick and skillful, he or she was thought to be quick, "in a smooth, easy-flowing way," whereas someone who was described as quick and clumsy was thought to be quick, "in a bustling way—the kind that . . . tips over the lamps" (Asch, 1946, p. 280).

Our results suggest that the subjectivity apparent even in covariation information allows highly related processes to operate during both impression formation and causal learning. The results reported here, along with phenomena such as order effects, are likely just a small set of the similarities between these two abilities. It remains unclear whether impression formation and causal learning rely on similar or identical processes, but the similarities between the literatures reviewed here suggest that it is a worthy avenue for future research.

Motivated Reasoning

Another relevant phenomenon within social psychology is motivated reasoning; people appear to evaluate arguments so as to protect their wishes, desires, and preferences (Pyszczynski, Greenberg, & Holt, 1985; Wyer & Frey, 1983). These evaluations seem quite similar to the interpretations observed in the current study with prior beliefs exerting an influence. Whereas some studies have suggested that self-relevance and threat, which are likely minimized in our paradigm, are critical for motivated reasoning to occur (Kunda, 1987, Experiments 3 and 4), others have found that biases could occur in nonthreatening contexts (Wason, 1960, 1966; Wisniewski & Medin, 1994). For example, Gilovich (1983) reported that, within people who gamble on football games, winners preferred to make attributions to the relative skill of the two teams, whereas losers preferred to make attributions to luck. Importantly, attributions differed only when the game's description included fortuitous plays (i.e., flukes) that could be used to justify the luck attribution. Thus, the differential evaluation occurred only when there was some degree of ambiguity (Team A lost, but there were anomalous events). As we argued above, covariation information necessarily includes ambiguity and is thus similarly ripe for differing attributions. Furthermore, the alternative cause present in Experiment 1C provided a powerful fluke and allowed our learners to maintain their existing beliefs.

The debate about how such biases operate remains generally unresolved. Our results suggest that similar processing is occurring in our causal learning task, albeit with relatively nonthreatening, nonpersonal influences. Thus, it seems useful to explore causal learning alongside motivated reason until there is evidence that these are qualitatively different.

Abductive and Scientific Reasoning

Last, we note that our results are also relevant for the study of scientific reasoning and abductive reasoning more generally. As has been noted (Kuhn & Dean, 2004), scientific conclusions rely on making appropriate causal inference based on covariation information. Most relevant to the current study, scientists must deal with data that contradict currently accepted theory. Chinn and Brewer (1998) outlined a taxonomy of eight possible responses to anomalous scientific data. Several of these responses are related to the interpretations we observed in the current results. First, Chinn and Brewer described the process of reinterpretation as one in which anomalous data are accepted as valid but explained in such a way as to allow the current scientific theory to remain unchanged. The process of rejection is one in which the anomalous data are simply deemed invalid (e.g., that particular experiment is inconsistent with my theory, but it was not well controlled) and thus do not alter the current scientific theory (see Fugelsang & Thompson, 2000, for related results). Ignorance is similar, except that there is no attempt to explain away the anomalous data; they are simply ignored.

These strategies for dealing with anomalous scientific data are highly related to the interpretations that our participants made when confronted with contradictory covariation information. They also further reinforce the critical but underexplored connection between processing of individual, token-level pieces of information of the sort Chinn and Brewer (1998) focused on and larger, more abstract theoretical beliefs. Our data also suggest that Chinn and Brewer's framework, which was developed to account for processing in the content-rich domain of scientific reasoning, can be fruitfully applied to reasoning in the significantly more austere circumstances present in traditional causal learning studies.

Conclusions

We have provided evidence that covariation information, the presumed basis of learning, contains a significant amount of ambiguity. This evidence stands in contrast to the assumptions embodied by current learning theories. The ambiguity present in covariation information allows learners to process such information in a flexible manner. In the current study, individual pieces of covariation information were interpreted differently depending on a variety of factors. These interpretations, in turn, influenced how learning progressed. We then utilized the observed pattern of interpretations to construct a sequence to demonstrate that purportedly positive covariation evidence could have negative influences. The linkage between the online processing of ambiguous covariation information and learning itself is relatively unexplored territory but suggests a set of processes common to both causal learning and causal reasoning. By further understanding how individual explanations are related to learning, researchers will have a richer understanding of causal cognition.

References

Anderson, N. H., & Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *Journal of Verbal Learning and Verbal Behavior, 2,* 379–391. doi:10.1016/S0022-5371(63)80039-0

Anderson, N. H., & Jacobson, A. (1965). Effect of stimulus inconsistency and discounting instructions in personality impression formation. *Journal of Personality and Social Psychology, 2,* 531–539. doi:10.1037/h0022484

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258–290. doi:10.1037/h0055756

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 238–249. doi:10.1037/0278-7393.31.2.238

Busemeyer, J. R. (1991). Intuitive statistical estimation. *Contributions to Information Integration Theory, 1,* 189–215.

Catena, A., Maldonado, A., Megias, J. L., & Frese, B. (2002). Judgement of frequency, belief revision, and serial processing of causal information. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 55*(B), 267–281. doi:10.1080/02724990244000007

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74,* 271–280. doi:10.1037/h0027592

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. doi:10.1037/0033-295X.104.2.367

Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science*

Teaching, 35, 623–654. doi:10.1002/(SICI)1098-2736(199808)35:6<623::AID-TEA3>3.0.CO;2-O

Collins, D. J., & Shanks, D. R. (2002). Momentary and integrative response strategies in causal judgments. *Memory & Cognition, 30,* 1138–1147. doi:10.3758/BF03194331

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology, 47,* 109–121. doi:10.1016/S0022-2496(02)00016-0

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 67–74). Cambridge, MA: MIT Press.

Danks, D., & Schwartz, S. (2005). Causal learning from biased sequences. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 542–547). Mahwah, NJ: Erlbaum.

Dennis, M. J., & Ahn, W. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition, 29,* 152–164. doi:10.3758/BF03195749

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3–19. doi:10.1037/0033-2909.99.1.3

Fugelsang, J. A., & Thompson, V. A. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology, 54,* 15–32. doi:10.1037/h0087327

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology, 44,* 1110–1126. doi:10.1037/0022-3514.44.6.1110

Glautier, S. (2008). Recency and primacy in causal judgments: Effects of probe question and context switch on latent inhibition and extinction. *Memory & Cognition, 36,* 1087–1093.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25,* 565–610. doi:10.1207/s15516709cog2504_3

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384. doi:10.1016/j.cogpsych.2005.05.004

Hendrick, C., & Costantini, A. (1970). Effects of varying trait inconsistency and response requirements on the primacy effect in impression formation. *Journal of Personality and Social Psychology, 15,* 158–164. doi:10.1037/h0029203

Jenkins, H. M., & Ward, W. C. (1965). Judgement of contingency between responses and outcomes. *Psychological Monographs, 79* (1, Whole No. 594), 1–17.

Kuhn, D., & Dean, D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development, 5,* 261–288. doi:10.1207/s15327647jcd0502_5

Kunda, Z. (1987). Motivation and inference: Self-serving generation and evaluation of evidence. *Journal of Personality and Social Psychology, 53,* 636–647. doi:10.1037/0022-3514.53.4.636

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55,* 232–257. doi:10.1016/j.cogpsych.2006.09.006

López, F. J., Shanks, D. R., Almaraz, J., & Fernández, P. (1998). Effects of trial order on contingency judgments: A comparison of associative and probabilistic contrast accounts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 672–694. doi:10.1037/0278-7393.24.3.672

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. (2008). Sequential causal learning in humans and rats. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 188–195). Washington, DC: Cognitive Science Society.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–984. doi:10.1037/a0013256

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science, 34,* 113–147. doi:10.1111/j.1551-6709.2009.01058.x

Luhmann, C. C., & Ahn, W. (2007). BUCKLE: A model of unobserved cause learning. *Psychological Review, 114,* 657–677. doi:10.1037/0033-295X.114.3.657

Marsh, J. K., & Ahn, W. (2006). Order effects in contingency learning: The role of task complexity. *Memory & Cognition, 34,* 568–576. doi:10.3758/BF03193580

Matute, H., Vegas, S., & De Marez, P. (2002). Flexible use of recent information in causal and predictive judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 714–725. doi:10.1037/0278-7393.28.4.714

Pyszczynski, T., Greenberg, J., & Holt, K. (1985). Maintaining consistency between self-serving beliefs and available data: A bias in information evaluation. *Personality and Social Psychology Bulletin, 11,* 179–190. doi:10.1177/0146167285112006

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

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110,* 101–120. doi:10.1037/0096-3445.110.1.101

Shanks, D. R., Holyoak, K. J., & Medin, D. L. (Eds.). (1996). *Causal learning.* San Diego, CA: Academic Press.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7,* 337–342. doi:10.1111/j.1467-9280.1996.tb00385.x

Stewart, R. H. (1965). Effect of continuous responding on the order effect in personality impression formation. *Journal of Personality and Social Psychology, 1,* 161–165. doi:10.1037/h0021641

Vandorpe, S., De Houwer, J., & Beckers, T. (2007). Outcome maximality and additivity training also influence cue competition in causal learning when learning involves many cues and events. *Quarterly Journal of Experimental Psychology, 60,* 356–368. doi:10.1080/17470210601002561

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25,* 127–151. doi:10.1006/lmot.1994.1008

Waldmann, M. R. (1996). Knowledge-based causal induction. *Psychology of Learning and Motivation, 34,* 47–88. doi:10.1016/S0079-7421(08)60558-7

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129–140. doi:10.1080/17470216008416717

Wason, P C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Harmondsworth, England: Penguin.

Wasserman, E. A., Kao, S. F., Van Hamme, L., Katagiri, M., & Young, M. E. (1996). Causation and association. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *Causal learning* (pp. 207–264). San Diego, CA: Academic Press.

White, P. A. (1995). Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory & Cognition, 23,* 243–254. doi:10.3758/BF03197225

White, P. A. (2002). Causal attribution from covariation information: The evidential evaluation model. *European Journal of Social Psychology, 32,* 667–684. doi:10.1002/ejsp.115

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18,* 221–281. doi:10.1207/s15516709cog1802_2

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136,* 82–111. doi:10.1037/0096-3445.136.1.82

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology, 47,* 276–332. doi:10.1016/S0010-0285(03)00036-7

Wyer, R. S., & Frey, D. (1983). The effects of feedback about self and others on the recall and judgments of feedback-relevant information. *Journal of Experimental Social Psychology, 19,* 540–559. doi:10.1016/0022-1031(83)90015-X

Yates, J. F., & Curley, S. (1986). Contingency judgment: Primacy effects and attention decrement. *Acta Psychologica, 62,* 293–302. doi:10.1016/0001-6918(86)90092-2

Appendix A

Rescorla–Wagner's Assumptions About the Influence of Covariation

The following assumes a single cause and the ever-present background (Rescorla & Wagner, 1972). We also assume sufficiently moderate learning rates (α_{bg}, α_{cause}, and $\beta \sim 0.5$) because extremely large learning rates will necessarily, but artificially, lead to unexpected, nontraditional behavior from the model. As is convention, we also assume that associative strengths (V) begin at zero. Here, again, the model will tend to exhibit unexpected behavior if the initial values of V are manually set to extreme initial values.

To achieve deviations from the traditional adjustments made by the Rescorla–Wagner model (RW), one needs to manipulate the quantity ($V_{\text{bg}} + V_{\text{cause}}$) to be less than 0 or greater than λ. It can be shown (using the derivation described by Danks, 2003, assuming $\lambda = 1$) that the equilibrium value of this sum is equal to $P(E|C)$. This necessarily bounds ($V_{\text{bg}} + V_{\text{cause}}$) to the range [0, 1] in the long run because probabilities cannot fall outside this range. When the sum ($V_{\text{bg}} + V_{\text{cause}}$) is within this range, all cells of the contingency matrix result in associative strength changes that match the traditional influences embodied by models such as ΔP and power PC. RW can, however, be made to achieve values of ($V_{\text{bg}} + V_{\text{cause}}$) that momentarily escape this range under certain circumstances. Below, we illustrate how situations can be constructed such that RW makes atypical, nontraditional adjustments given each of the covariation types under discussion in the current study (Cells A and B). As we demonstrate, deviations can only occur under specialized circumstances, typically requiring a specific sequence of observations. Most importantly, these are not the sequences used in the current experiments.

Cell A

Upon encountering an observation from Cell A of the covariation matrix, the equation for RW is

$$\Delta V = \alpha\beta(1 - \Sigma V),$$

assuming $\lambda = 1$, as in typical simulations of RW when the effect is present (though the following arguments also hold for other positive values of λ). In response to observations from Cell A, RW typically increases associative strengths because usually $\Sigma V < 1$. If, however, $\Sigma V > 1$, that is, $V_{\text{bg}} + V_{\text{cause}} > 1$, then Cell A observations will act to decrease associative strengths. Increasing both V_{bg} and V_{cause} until their sum exceeds 1 is extremely difficult. For instance, attempts to simply increase both V_{bg} and V_{cause} by presenting Cell A observations will not achieve $V_{\text{bg}} + V_{\text{cause}} > 1$ because RW builds in competition between causes (i.e., the cause and the background in this case). This means that repeated presentations of Cell A observations will cause the sum $V_{\text{bg}} + V_{\text{cause}}$ to asymptote to 1 (i.e., λ), rather than exceeding 1.

One can, instead, attempt to increase V_{bg} and V_{cause} separately. Increasing V_{bg} can be done without increasing V_{cause} by presenting Cell C observations. However, to the extent that V_{bg} is strengthened, subsequent attempts to increase only V_{cause} without increasing V_{bg} will fail because causes are always presented in the presence of the background, and the two cues compete, keeping $V_{\text{bg}} + V_{\text{cause}} < 1$.

Thus, V_{cause} must first be increased to some value greater than zero by presenting Cell A observations. At that point, V_{bg} may then be increased without increasing V_{cause} by presenting Cell C observations until $V_{\text{bg}} + V_{\text{cause}} > 1$. Once this has been done, Cell A observations will produce a negative prediction error ($\Delta V < 0$). The order of this sequence is important because, as explained above, any attempt to first increase V_{bg} without increasing V_{cause} will limit the degree to which one can subsequently increase V_{cause}. Note that this is not the order used in the current experiments or in typical causal learning experiments.

In the current study, Cell A observations during the second half of our negative–positive sequences cannot result in negative prediction errors because, at the end of the relatively homogeneous negative block consisting mostly of Cells B and C, V_{cause} will have neared its asymptote of -1 (due to Cell B observations) and V_{bg} will have neared its asymptote of 1 (due to Cell C observations). As a result, their sum (ΣV) will be near 0. (ΣV will be no less than 0 because V_{cause} can only be adjusted downward after V_{bg} is already positive, so V_{bg} has a head start; yet ΣV will be much less than 1). Thus, upon encountering Cell A observations in the second, positive block, $V_{\text{bg}} + V_{\text{cause}} << 1$, resulting in large, positive prediction errors, in contrast to the negative and neutral interpretations found in the current study. (In case a reader wonders about the role of a few Cells A and D presented in the negative block, note that the minority of D observations in the negative block only act to decrease V_{bg}. The minority of A cells increase both V_{bg} and V_{cause} by the same magnitude but only, at most, to the extent needed to achieve $V_{\text{bg}} + V_{\text{cause}} = 1$, never such

(*Appendices continue*)

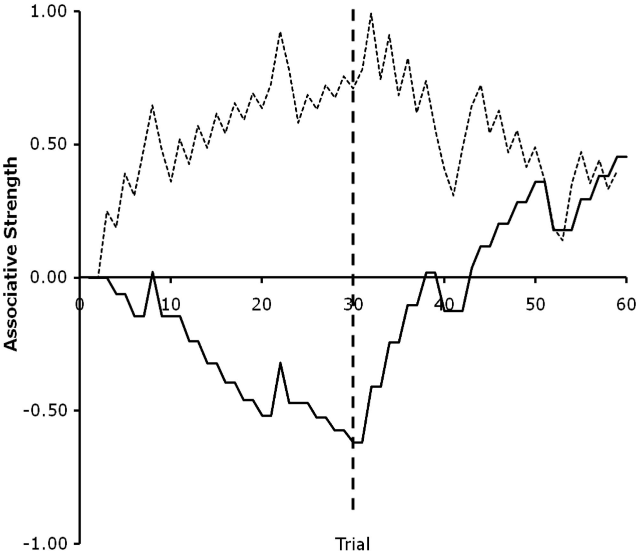


Figure A1. Simulation results achieved by applying the Rescorla–Wagner model to a representative negative–positive sequence used in the current studies. The solid line represents the associative strength of the cause (i.e., V_{cause}). The broken line represents the associative strength of the constant experimental background (i.e., V_{bg}). The vertical line represents the transition from the first, negative half of the sequence to the second, positive half of the sequence. In this simulation ($\lambda = 1$, α_{bg}, α_{cause}, and $\beta = 0.5$), the sum of $V_{\text{bg}} + V_{\text{cause}}$ at the conclusion of the first half of the sequence is .16, which means that subsequent observations from Cell A will act to increase associative strengths, as expected.

that $V_{\text{bg}} + V_{\text{cause}} > 1$.) Figure A1 shows a simulation result of a representative negative–positive sequence used in Experiment 1A.

Cell B

Upon encountering an observation from Cell B of the covariation matrix, the equation for RW is

$$\Delta V = \alpha\beta(0 - \Sigma V)$$

because $\lambda = 0$ when the outcome is absent, and RW typically decreases associative strengths because $\Sigma V > 0$. If $\Sigma V < 0$, then $\Delta V > 0$ and Cell B observations will act to increase associative strengths.

Again, achieving Σ is extremely difficult. For instance, one may think that $\Sigma V < 0$ can be achieved by decreasing either V_{bg} and V_{cause}, by using Cell B or Cell D. However, since V_{bg} and V_{cause} are initially zero, these observations result in $\Delta V = \alpha\beta(0 - 0) = 0$ and thus have no influence. To get around this situation, we can manufacture a trial sequence such that ΣV first takes a nonzero value and then decrease V_{bg} and V_{cause} to be less than zero. The simplest sequence that achieves this appears to be a Cell C observation, followed by the Cell B observation, followed by Cell D observations. Under normal circumstances, Cell C observation followed by Cell B observation will keep $V_{\text{bg}} > -V_{\text{cause}}$ (and thus $V_{\text{bg}} + V_{\text{cause}} > 0$). This is because the initial presentation of the Cell C observation gives V_{bg} a head start over V_{cause}, which is required to allow the subsequent Cell B observation to decrease V_{cause}. Thus, once $V_{\text{cause}} < 0$, one needs to make it such that $\Sigma V < 0$ by presenting Cell D observations until $V_{\text{bg}} < -V_{\text{cause}}$. Cell B observations will now produce a positive prediction error. Again, this exercise is intended to illustrate just how difficult it is to achieve a positive prediction error from Cell B observations using RW and to further demonstrate that the required sequence is not used in the current experiment or in typical causal induction experiments.

Figure A2 shows simulation results of RW using a sample sequence for our positive–negative order condition. Cell B observations during the second half of our positive–negative sequences cannot result in positive prediction errors because, at the end of the relatively homogeneous positive block, V_{bg} will have neared its asymptote of 0, V_{cause} will be near its asymptote of λ, and their sum will be greater than 0. Upon encountering Cell B observations, these observations will generate large, negative prediction errors.

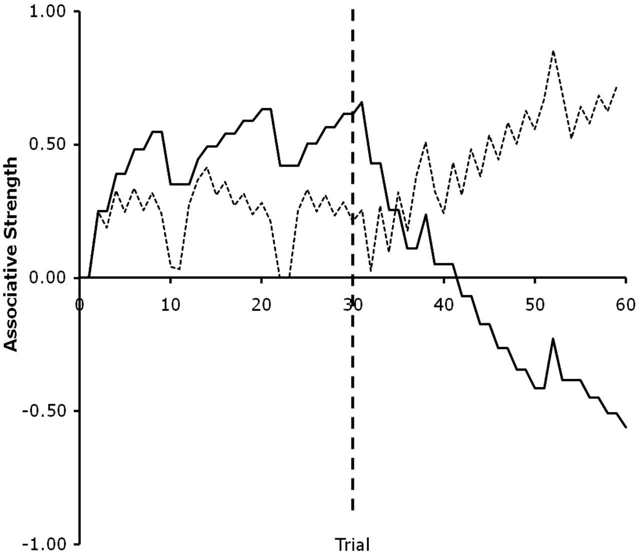


Figure A2. Simulation results achieved by applying the Rescorla–Wagner model to a representative positive–negative sequence used in the current studies. The solid line represents the associative strength of the cause (i.e., V_{cause}). The broken line represents the associative strength of the constant experimental background (i.e., V_{bg}). The vertical line represents the transition from the first, positive half of the sequence to the second, negative half of the sequence. In this simulation ($\lambda = 1$, α_{bg}, α_{cause}, and $\beta = 0.5$), the sum of $V_{\text{bg}} + V_{\text{cause}}$ at the conclusion of the first half of the sequence is .91, which means that subsequent observations from Cell B will act to decrease associative strengths, as expected.

(*Appendices continue*)

Appendix B

Calculating Composite Scores

To compute the composite interpretation and causal strength scores we used in the analysis of Experiment 1, we first reversed for all measures elicited in the negative–positive order condition (i.e., −1 to 1, and 1 to −1) and averaged these scores with the unmodified scores from the positive–negative order (e.g., see Figure B1 under “Coding (reversed)” and “Average” in the negative–positive order condition). This average measured the extent to which participants’ interpretations were consistent with the first block of the sequence because 1 represents a positive interpretation for the positive–negative order and a negative interpretation for the negative–positive order, whereas −1 represents the opposite.

More specifically, the left side of Figure B1 illustrates a case in which a participant showed a full-blown neighborhood effect in the first block. In this case, the participant would have provided only positive interpretations in the first block from the positive–negative order (i.e., coded as all 1s) and also only negative interpretations in the first block from the negative–positive order (i.e., reverse-coded as all 1s). The resulting first-block composite score would thus be (1 + 1 + 1 + 1)/4 = 1.

The left side of Figure B2 illustrates a case in which a participant instead provided first-block interpretations that were always positive for Cell A and always negative for Cell B. The average composite scores in this case are [1 + 1 + (−1) + (−1)]/4 = 0.

Similarly, we computed the composite scores for the second block (i.e., reverse-coding the negative–positive order condition and averaging its second-block scores with the second-block scores from the positive–negative order). This time, higher composite scores imply an expectancy bias (second-block interpretations that were consistent with the covariation information presented in the first block). The right side of Figure B1 illustrates the case in which a participant exhibited a full-blown expectancy bias, making interpretations entirely consistent with the first half of the sequence. In this case, the participant would give all positive interpretations in the second block of the positive–negative order (coded as 1s) and all negative interpretations in the second block of the negative–positive order (reverse-coded to be 1s), resulting in the composite score of 1. If a participant instead provided interpretations that were consistent with the traditional influence of covariation information (right side of Figure B2), this score would be 0.

Positive-Negative Order

	First Block (positive)				Second Block (negative)			
Cell Type	A	A	B	B	A	A	B	B
Response	Pos	Pos	Pos	Pos	Pos	Pos	Pos	Pos
Coding	1	1	1	1	1	1	1	1

Negative-Positive Order

	First Block (negative)				Second Block (positive)			
Cell Type	A	A	B	B	A	A	B	B
Response	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg
Coding (reversed)	1	1	1	1	1	1	1	1
Average	1	1	1	1	1	1	1	1

Figure B1. Hypothetical data of a participant showing a full-blown neighborhood effect in the first block and a full-blown expectancy-based effect in the second block.

Positive-Negative Order

	First Block (positive)				Second Block (negative)			
Cell Type	A	A	B	B	A	A	B	B
Response	Pos	Pos	Neg	Neg	Pos	Pos	Neg	Neg
Coding	1	1	-1	-1	1	1	-1	-1

Negative-Positive Order

	First Block (negative)				Second Block (positive)			
Cell Type	A	A	B	B	A	A	B	B
Response	Pos	Pos	Neg	Neg	Pos	Pos	Neg	Neg
Coding (reversed)	1	1	-1	-1	1	1	-1	-1

Figure B2. Hypothetical data of a participant showing no neighborhood effect in the first block and no expectancy-based effect in the second block.

Finally, we obtained composite causal strength scores by subtracting causal strength judgments in the negative–positive condition from those in the positive–negative condition and dividing by 2: [(positive–negative judgment – negative–positive judgment)/2]. This composite score again measured the extent to which judgments were consistent with the first block. For instance, if causal strength judgments were based on only the first half of the sequence, then it would be coded as 100 in the positive–negative condition and as −100 in the negative–positive condition, resulting in a composite score of [100 − (−100)]/2 = 100. Thus, learners with composite causal strength scores near 100 would be demonstrating greater primacy effects. In contrast, if causal strength judgments were based entirely on the second half of the sequence, then it would be coded as −100 in the positive–negative condition and as 100 in the negative–positive condition, resulting in a composite score of [−100 − (100)]/2 = −100. Finally, if their causal strength was not biased in favor of either of the blocks, composite causal strength scores would be 0.

Received March 16, 2009
Revision received September 18, 2010
Accepted September 23, 2010 ■