American Psychological Association

Journal of Experimental Psychology: General

0096-3445/18/$12.00

2018, Vol. 147, No. 4, 485–513
http://dx.doi.org/10.1037/xge0000423

Causal Strength Induction From Time Series Data

Kevin W. Soo and Benjamin M. Rottman
University of Pittsburgh

One challenge when inferring the strength of cause-effect relations from time series data is that the cause and/or effect can exhibit temporal trends. If temporal trends are not accounted for, a learner could infer that a causal relation exists when it does not, or even infer that there is a positive causal relation when the relation is negative, or vice versa. We propose that learners use a simple heuristic to control for temporal trends—that they focus not on the *states* of the cause and effect at a given instant, but on how the cause and effect change from one observation to the next, which we call *transitions*. Six experiments were conducted to understand how people infer causal strength from time series data. We found that participants indeed use transitions in addition to states, which helps them to reach more accurate causal judgments (Experiments 1A and 1B). Participants use transitions more when the stimuli are presented in a naturalistic visual format than a numerical format (Experiment 2), and the effect of transitions is not driven by primacy or recency effects (Experiment 3). Finally, we found that participants primarily use the direction in which variables change rather than the magnitude of the change for estimating causal strength (Experiments 4 and 5). Collectively, these studies provide evidence that people often use a simple yet effective heuristic for inferring causal strength from time series data.

Keywords: causal learning, covariation detection, time series, temporal trend

Inferring the causal influence one variable has on another in a time series setting is a complex task. The fundamental problem, detailed by Yule (1926), is that the two variables can undergo temporal trends, otherwise known as "secular trends." Temporal trends can make it appear as if there is a positive or negative relationship between two variables even when there is no direct relation, such as the famous example that ice cream sales are correlated with drownings (because of changes in weather). Temporal trends can also make a negative causal relation appear positive, or vice versa. For example, the U.S. economy and the price of oil have generally increased over time (a positive correlation), even though increases in the price of oil cause the economy to contract on a smaller time scale. The main question of the current research is whether and how laypeople are able to "control for" temporal trends and reach accurate conclusions about the strength of a cause-effect relation.

The ability to learn about the strength of a relation between two probabilistic variables—covariation detection—is a fundamental cognitive function, and underlies reasoning across many areas of psychology. Covariation detection is a critical mechanism for categorization (Kutzner & Fiedler, 2015; Vogel, Kutzner, Freytag, & Fiedler, 2014) and stereotype formation (Le Pelley et al., 2010; Sherman et al., 2009). The disruption of this ability—for example, perceiving correlation when there is none (illusory correlation)—has been found to play a role in a range of clinical disorders like depression (Alloy & Abramson, 1979), phobias (Ohman & Mineka, 2001), and schizophrenia (Balzan, Delfabbro, Galletly, & Woodward, 2013; Díez-Alegría, Vázquez, & Hernández-Lloreda, 2008; Huq, Garety, & Hemsley, 1988).

Past research on *covariation detection* and *causal strength induction*[1] has primarily focused on settings in which the observations of the variables are temporally independent, and thus there are no temporal trends. For example, the cover stories could involve 20 hypothetical patients, and participants are tasked with assessing whether patients who take a medication are more or less likely to have a headache relative to patients who do not take the medication. Less research has focused on covariation detection or causal strength induction from time series data—when variables are observed over time and can exhibit temporal trends. We believe these cases involving time series data are especially compelling; they are often the sorts of examples used to convey the point that correlation does not imply causation due to "spurious correlations" (e.g., the notorious ice cream and drownings example; see Vigen, 2015, for many more examples). Indeed, correlations due to temporal trends were the cases that Yule (1926) was concerned about.

Kevin W. Soo and Benjamin M. Rottman, Department of Psychology, University of Pittsburgh.

This research was supported by NSF 1430439. Experiments 1A, 2, and 4 were presented at the 38th Annual Conference of the Cognitive Science Society (Soo & Rottman, 2016). Raw data and stimuli for the experiments presented here are available on the Open Science Framework: osf.io/9avdm/.

Correspondence concerning this article should be addressed to Kevin W. Soo, Department of Psychology, University of Pittsburgh, LRDC 720, 3939 O'Hara Street, Pittsburgh, PA 15260. E-mail: kevin.soo@pitt.edu

[1] We use *covariation detection* to refer to the task of inferring the strength between two variables, neither of which is proposed or assumed a priori to be the cause of the other. We use *causal strength induction* to refer to the task of inferring the strength of one variable that is presumed to be a cause on another variable that is presumed to be an effect. Our studies focus on causal strength induction, though these two tasks are highly similar in many respects.

To give a concrete example of the subject of this research, Figure 1A plots two variables over time. X is stipulated to be a potential cause of Y. Overall, the two variables are positively correlated because they both increase over time. How does one infer the strength of the potential causal influence of X on Y? One aspect of the data reveals a clue: From one observation to the next, when X increases, Y usually decreases, and when X decreases, Y usually increases. This can be seen clearly in Figure 1B, which plots the *difference scores* at each time point for both variables—the amount they changed from the last time point. We argue that these local patterns provide a way to normatively estimate the causal effect of X on Y, and that people in fact use these local patterns for inferring causal strength, in addition to the global correlation.

The outline of the introduction is as follows. First, we discuss previous research on causal strength learning in atemporal settings. Following this, we discuss work on causal learning in time series settings. Finally, we propose a simple heuristic that people may use to control for temporal trends when inferring causal strength in time series settings. In essence, the heuristic we propose is that instead of calculating the correlation between the cause and effect, that people (roughly) calculate the correlation between the change in the cause and the change in the effect from one observation to the next. We also discuss one factor that may moderate the use of this heuristic, and we propose two different versions of this heuristic.

Causal Learning in Atemporal, Cross-Sectional Settings

Although causal strength induction has been studied heavily, most existing theories do not apply to the question addressed in the current research because they typically involve scenarios that are atemporal. The cover stories often resemble randomized controlled experiments, for example, estimating the efficacy of a medication by comparing the percent of patients who experience a symptom across two groups of patients, one of which received a medication and the other of which did not. We refer to this as the *cross-sectional* paradigm, analogous to cross-sectional research designs, to convey that the data are from a single cross-section of time, not from a time series.

The bulk of research on causal strength induction has involved a cross-sectional paradigm with a binary cause and a binary effect. In such experiments, the normative theories for judging causal strength involve a mental calculation somewhat similar to a chi-square test of association (Buehner, Cheng, & Clifford, 2003; Cheng, 1997; Cheng & Novick, 1992; Griffiths & Tenenbaum, 2005; for reviews, see Hattori & Oaksford, 2007; Holyoak & Cheng, 2011; Perales & Shanks, 2007). A less common paradigm involves a binary cause and a continuous effect, in which case the *t* test has been proposed as the normative model (Obrecht, Chapman, & Gelman, 2007; Saito, 2015).

There has been considerably less research on causal strength induction in cross-sectional settings when the cause and effect both fall on a continuous or multilevel scale, as they do in our experiments. The obvious contender for a normative model is Pearson's correlation coefficient *r*, which has long been used both as a normative and descriptive model of covariation detection (Beach & Scopp, 1966; Crocker, 1981; Erlick, 1966; Erlick & Mills, 1967; Lane, Anderson, & Kellam, 1985). However, one potential concern with using *r* as a normative model for causal strength is that it is symmetric; it does not distinguish between X or Y being the cause versus effect. In causal learning paradigms with binary causes and effects, people treat the cause and effect differently when inferring causal strength, and symmetric measures like chi-square have never gained traction as models of causal strength (Cheng, 1997; Griffiths & Tenenbaum, 2005). However, because so far no other model of causal strength has been proposed when there is a continuous cause and a continuous effect, we treat *r* as a reasonable default normative model.

The more pressing concern with using *r* as a normative model is that *r* is appropriate when the data are independent, such as in cross-sectional settings, but not in time series settings in which variables can exhibit temporal trends. This problem can be conceptualized as time being a confound of both X and Y (see Figure 2), in which case the appropriate statistical approach would be to control for time, for example, by computing a regression of Y ~

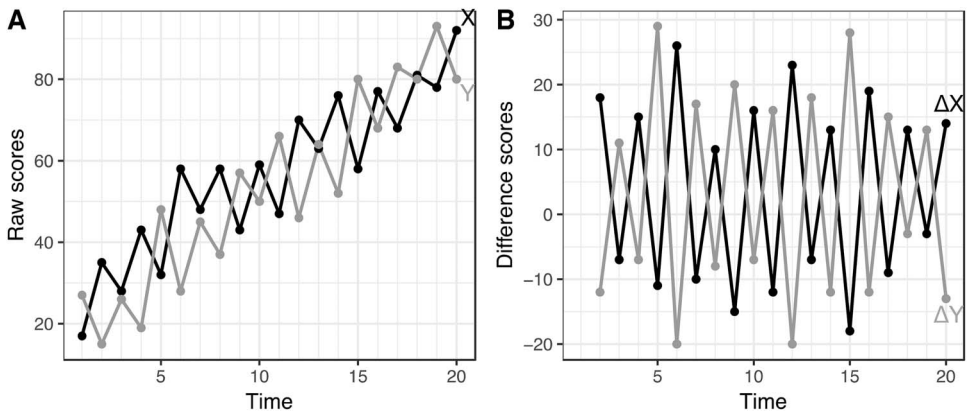

Figure 1. (A) Example time series data of a cause (X) and effect (Y). The raw scores of X and Y are displayed. (B) The difference scores, ΔX and ΔY. There are 19 difference scores for the 20 raw scores. Note the different vertical axes.

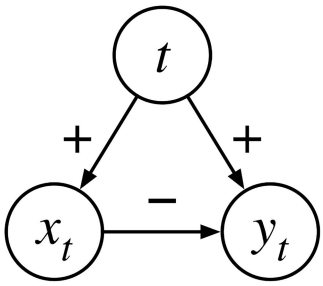

Figure 2. A causal graph depicting a potential relationship between X and Y when there is a temporal trend (*t*) influencing both over time. X negatively influences Y, but both X and Y increase over time.

X + *t*. (Calculating partial variance explained by X in the prior regression equation is also equivalent to calculating the "partial" correlation of X and Y; the partial correlation is the correlation between the residuals of X and the residuals of Y after partialing out *t*.) Our research investigates if and how learners account for time in these sorts of situations. The following section reviews previous research on causal learning in time series settings in order to build intuitions for the model we propose.

Causal Learning in Time Series Settings

There have been four efforts we consider most relevant for understanding causal learning in time series settings.

Causal Learning in Time Series Settings With Delay

The first set of research focused on how people learn about causal relationships from temporal streams of data in which one variable (the cause) occurs before another variable (the effect). One important finding is that people use temporal precedence—which variable occurs first—for determining which variable is the cause and which is the effect (Burns & McCormack, 2009; Lagnado & Sloman, 2004, 2006; McCormack, Frosch, Patrick, & Lagnado, 2015). Another related finding involves paradigms in which the goal is to infer causal strength, but there is a delay between the cause and the effect (Buehner & May, 2003, 2009; Greville & Buehner, 2010; Lagnado & Speekenbrink, 2010; Shanks, Pearson, & Dickinson, 1989). If learners do not have strong beliefs about the delay, a delay can lead to weaker causal strength judgments. However, if learners have strong expectations for the delay and the length of delay is consistent, they can parse the event stream so that the cause at one time point is associated with the effect at a subsequent time point. Parsing the data allows them, in theory, to apply the same models developed for cross-sectional scenarios to the parsed data (Buehner, 2005; Greville, Cassar, Johansen, & Buehner, 2013; Hagmayer & Waldmann, 2002).

Causal Learning With Nonstationary Time Series Data

The most critical feature of time series data in the current experiments is that the data can exhibit *nonstationary trends* (e.g., Figure 1A). For time series variables on a continuous scale, a nonstationary trend means that the average value of a variable changes over time; this average value can be calculated with a moving average within some window of time. Binary time series variables can also exhibit trends, for example, when a variable tends to remain in the same state for multiple observations in a row rather than switching states randomly, which is typically called positive *autocorrelation*, though could also be considered a type of nonstationarity.[2] The following three sets of research have focused on causal learning in nonstationary or autocorrelated time series environments.

One set of research examined whether people can learn the direction of a causal relation (which variable is the cause and which is the effect) from time series data when there is no delay between the cause and the effect. Suppose that X is the cause and Y is the effect. In a time series setting, when X increases or decreases from one observation to the next, Y would also be expected to increase or decrease. However, sometimes Y might change due to the influence of unobserved factors. When Y changes due to the influence of an unobserved factor, if X is positively autocorrelated, X will tend to remain fairly stable or follow the preexisting trend. This means that if one variable (Y) changes but another (X) does not, the variable that changed on its own (Y) is more likely to be the effect. Both adults and children notice this asymmetry in how X and Y change and use it to infer that X causes Y (Rottman & Keil, 2012; Rottman, Kominsky, & Keil, 2014; Soo & Rottman, 2014). The important theoretical innovation here was to uncover that people utilize the changes or what we call the *transitions* in the variables for causal inference.

Two other studies have examined how people infer causal strength, rather than structure, in time series settings. White's (2015) participants were asked to judge the influence of a chemical injected into a patient's bloodstream on the concentration of blood cells. The patient's blood cells were tracked hourly over 24 hr and the injection occurred during one particular hour, so the data comprised an "interrupted time series" design. The blood cell concentration tended to increase across the 24 hr, though the features of the trend (e.g., the apex, when the increase started relative to the injection) were varied. One factor that determined participants' judgments of the strength of the injected chemical was the difference in the blood cell count after versus before the injection, similar to how statisticians calculate difference scores for prepost study designs. However, another finding suggested that the participants had difficulty fully accounting for important aspects of the time series data. In White's (2015) Experiment 1A, participants judged the injection as causing an increase in the blood cell count even if the blood cell count had already started increasing before the injection. Those judgments can be interpreted as participants failing to control for nonstationarity in the time series.

Rottman (2016) investigated another causal strength learning task. Participants had to learn which of two medications produced a bigger decrease in pain. In this cover story, participants chose to try either Medication 1 or Medication 2 each day (trial) for 14 simulated days. At the end of each day a pain score was revealed.

[2] Technically, nonstationarity in the generative process of a time series causes it to be autocorrelated across multiple time lags, though a variable could be autocorrelated without being non-stationary (Shumway & Stoffer, 2011).

Critically, the baseline amount of pain was nonstationary—it fluctuated in unpredictable waves—which meant that comparing Medication 1 for Days 1–7 versus Medication 2 for Days 8–14 could produce a poor comparison if the baseline amount of pain tended to increase or decrease across the 14 days. Rottman (2016) found that participants used at least two strategies for estimating the difference in the effectiveness of the two medicines. The first strategy was to simply take the means of the pain scores when Medicine 1 was tried minus the means of the pain scores when Medicine 2 was tried, similar to a *t* test. The second strategy involved first computing the change in the pain from one day to the next, and then comparing the average change in the pain score on days when Medicine 1 was tried versus days when Medicine 2 was tried.

In sum, the studies cited in this section suggest that in time series settings, people make use of how the cause and effect change from one observation to the next. The current research further tests how people infer causal strength in time series settings and makes a number of contributions beyond the work of Rottman (2016). First, in that study, participants were in control of the cause, meaning the process of judging causal strength was confounded with the process of choosing which cause to test on each trial. The current experiments were designed specifically to study the causal strength judgment process on its own. Second, whereas Rottman (2016) studied cases with a binary cause and a continuous effect, the current experiments studied how participants learn about the causal relations between a continuous cause and a continuous effect, which has typically been neglected in the causal learning literature. Third, the experiments presented here study the compelling situation in which controlling for a temporal trend would lead to the opposite conclusion as not controlling for the trend (e.g., Figure 1).

A Process for Estimating Causal Strength From Transitions

In the following sections, we propose and offer justifications for a process that learners may use to control for temporal trends in time series data, discuss a possible moderator of this process, and present two alternative versions of this process.

Normative, Perceptual, and Cognitive Justifications

Statistics. In time series analysis, a standard procedure for accounting for nonstationarity is to compute a difference score on the variables with trends before conducting other analyses. Taking a difference accounts for linear trends, and taking a difference of differences accounts for quadratic trends. One benefit of accounting for temporal trends by taking difference scores rather than regressing out time is that no parameters need to be estimated (Shumway & Stoffer, 2011). We propose that when judging causal strength from time series data, that people, roughly, compute a correlation between the difference scores for X (ΔX) and the difference scores for Y (ΔY). We call this model $r_{\Delta\text{Continuous}}$; the label "continuous" will be explained below as we discuss the different ways transitions can be encoded.

In Appendix A, we provide two proofs of the usefulness of $r_{\Delta\text{Continuous}}$. First, in Part 1, we prove that $r_{\Delta\text{Continuous}}$ uncovers the true causal influence of X on Y when there is a temporal confound (as in Figure 1); it partials out a linear temporal trend. Second, in Part 2, we prove that in stationary environments (no temporal trend) with large samples, $r_{\Delta\text{Continuous}}$ equals the correlation between the raw states of X and Y, which we call r_{States}. Furthermore, we ran a simulation to measure the precision of $r_{\Delta\text{Continuous}}$ compared to r_{States}. We randomly generated stationary data sets in which $r^2 \approx .50$, and calculated the standard deviations of r_{States} and $r_{\Delta\text{Continuous}}$. Though the standard deviation for $r_{\Delta\text{Continuous}}$ was a bit higher for small data sets, it converged to the standard deviation of r_{States} for large data sets. This implies that $r_{\Delta\text{Continuous}}$ is a fairly good approximation of the true causal strength in both stationary and nonstationary environments. In contrast, r_{States} is only useful in stationary environments because in nonstationary environments it can grossly misestimate the causal strength.

Psychology. Aside from the normative justification that people should focus on changes or transitions instead of states when inferring causal strength in time series settings, there is evidence that people do focus on changes in stimuli. Evidence from perception and psychophysics suggests that people are poor at accurately judging absolute levels of perceptual stimuli (e.g., size, weight, pain, etc.; see Brown, Marley, Donkin, & Heathcote, 2008; Donkin, Rae, Heathcote, & Brown, 2015). One reason is perceptual adaptation to a stimulus (Helson, 1948, 1964; Restle & Merr, 1968; Sarris, 1967); after someone adapts to the level of a stimulus, the stimulus loses focus, and regains focus when the stimulus changes. Another reason is the susceptibility of people's internal scales for representing quantities to context effects; observing one stimulus can influence the perceived magnitude of another (Frederick & Mochon, 2012; Krantz & Campbell, 1961). Likewise, when a stimulus changes, the new level of the stimulus can be perceived within the context of the prior stimulus, focusing attention on the amount of change rather than the absolute level of the stimulus. Consequently, it may be the norm to make perceptual judgments based on changes in levels rather than absolute levels (Thurstone, 1927a, 1927b).

Comparative encoding is not limited to low-level perceptual phenomena. Stewart, Brown, and Chater (2005) demonstrated that changes in a stimulus from one presentation to the next influence categorization decisions. Stewart, Chater, and Brown (2006) have proposed that people use comparisons in memory when making decisions of value, in temporal discounting, and in other economic decisions. Furthermore, theories of associative learning have been developed that learn by associating the onsets and offsets (transitions) of the conditioned and unconditioned stimuli instead of their presence versus absence (Klopf, 1988).

In sum, there are a number of statistical and psychological reasons for attending to changes, and one of the goals of this research is to assess whether, when, and how people do so.

A Moderator of Learning From Transitions

Unlike nonhuman animals, people have the unique ability to represent quantities symbolically, allowing precise reasoning with quantities (Feigenson, Dehaene, & Spelke, 2004). This enables people to distinguish not just between two and three, but also between 82 and 83. How does a numerical versus perceptual presentation of variables affect the causal learning process?

One the one hand, it is likely that an effect of transitions will be smaller when variables are presented numerically rather than per-

ceptually. Unlike perceptual stimuli, numbers allow learners to identify the absolute levels of variables, making judgments less susceptible to context effects. Such a finding would be interesting because focusing less on transitions would result in worse performance in a causal learning task with temporal trends, even though numerical symbols are typically viewed as facilitating learning and reasoning in ways that are impossible without numbers.

On the other hand, a numerical presentation format could instigate a more explicit form of reasoning than a perceptual presentation, and learners who are aware of obvious increasing or decreasing trends in time series data may account for these trends by focusing on transitions rather than states. In an extreme case, learners who intentionally control for trends may focus on transitions and ignore states even more than in the perceptual condition.

In Experiment 2, we test whether a symbolic (numerical) versus perceptual (visual) stimuli format moderates the use of transitions.

How are Transitions Encoded?

The model proposed thus far for how causal strength is computed is $r_{\Delta\text{Continuous}}$, a correlation between the changes in X and the changes in Y. A weakness of this model is that *r* (and linear regression models more generally, e.g., Brehmer, 1994) is a computational-level theory that fails to explain how the learner processes information in a tractable way (Marr, 1982). For this reason, we considered one simplified alternative of how people may use transitions for inferring causal strength.

We propose that people may simplify continuous variables by mentally discretizing them into a binary representation, making it easier to summarize the values of X and Y for computing causal strength (Marsh & Ahn, 2009). Discretizing is perhaps a dubious proposal for a model based on states, because discretization would require choosing an arbitrary cutoff such as 50 on the 0–100 scale. However, if participants focus on transitions rather than states, the transitions have natural cutoffs based on whether the variable increased (coded as +1), decreased (−1), or stayed the same (0) from the previous time point, which we refer to as *binary* difference scores. (Although technically this coding scheme has three values, +1, 0, and −1, when there are 100 possible states, it is very unlikely for a variable to stay at exactly the same state from one trial to the next, so essentially the coding is only +1 or −1, which is why we use the term binary).

Computing (roughly) a correlation between the binary difference scores of X and Y, $r_{\Delta\text{Binary}}$, only requires keeping track of four tallies, raising the possibility of discretization as a plausible heuristic that people may rely on to simplify causal learning in longitudinal scenarios. This proposal is similar to a theory by Stewart et al. (2006) that people only use ordinal comparisons (greater than, less than, or equal to) for many economic judgments.

As will be discussed in the designs of Experiments 4 and 5, $r_{\Delta\text{Continuous}}$ and $r_{\Delta\text{Binary}}$ are often fairly strongly correlated, making them difficult to separate empirically. However, this correlation means that just as $r_{\Delta\text{Continuous}}$ controls for temporal trends, so too does $r_{\Delta\text{Binary}}$. Furthermore, in Appendix A (Part 2), just as $r_{\Delta\text{Continuous}}$ approximates r_{States} in stationary environments, so too does $r_{\Delta\text{Binary}}$. Specifically, through simulations we show that although $r_{\Delta\text{Binary}}$ is not linearly related to r_{States} like $r_{\Delta\text{Continuous}}$ is, it is monotonically related to r_{States}. The precision of $r_{\Delta\text{Binary}}$ as an estimator of r_{States} is also worse than $r_{\Delta\text{Continuous}}$, but not all that bad, and improves with larger sample sizes. In sum, $r_{\Delta\text{Binary}}$ accomplishes many of the same functions as $r_{\Delta\text{Continuous}}$ but is simpler to compute, raising the possibility that people might use it as a heuristic, potentially in both stationary and nonstationary environments.

Experiments 4 and 5 were designed to test whether people use the magnitudes of changes ($r_{\Delta\text{Continuous}}$) or just the directions of the changes ($r_{\Delta\text{Binary}}$) for inferring causal strength.

Summary of Experiments

Experiments 1A and 1B tested the basic phenomenon of whether people use transitions for inferring causal strength. Experiment 2 tested whether the effect of transitions is moderated by using symbolic (numerical) versus perceptual (visual) stimuli. Experiment 3 ruled out a possible alternative explanation for the effects in Experiments 1 and 2, that participants inferred causal strength from a limited memory of observations (i.e., primacy/recency effects), rather than the transitions. Experiments 4 and 5 tested whether people encode transitions as magnitudes or discretely (as increases vs. decreases). All the reported experiments were approved by the University of Pittsburgh Human Research Protection Office.

Experiments 1A and 1B: Learning From States Versus Transitions

Figure 3 summarizes the way we tested whether people use transitions for inferring causal strength in time series settings in Experiments 1 and 2. Figure 3 shows the same 20 data points rearranged in three different orders: negative transitions, random, and positive transitions. Because all three data sets in Figure 3 have the same 20 observations of X and Y, taking the correlation of the states of X and Y results in a correlation of $r_{\text{States}} = .70$ for all three. Figure 3A shows the 20 data points in time series presentations with time on the abscissa, and the value of the cause and effect (X and Y) as the two lines on the ordinate. Figure 3B shows a scatterplot of the 20 data points with X on the abscissa and Y on the ordinate. This presentation makes it easy to see that these are the same 20 data points presented in different orders. The order of the observations is displayed with numbers and lines connecting the sequential data points.

The negative transitions panels in Figure 3 represent data that could arise when X and Y both increase over time due to an unobserved temporal confound but there is a negative causal influence of X on Y; it is the same data from Figure 1. This negative influence can be seen by examining how X and Y change from one observation to the next. In the time series presentation (Figure 3A), when X increases, Y decreases, and vice versa. This effect can also be seen in the corresponding scatterplot (Figure 3B): The lines connecting one observation to the next go from the top left to bottom right, and vice versa (negative transitions). Even though the correlation between X and Y is $r_{\text{States}} = .70$, $r_{\Delta\text{Continuous}} = -.97$.

In contrast, in the positive transitions panels in Figure 3, the order of the observations were rearranged such that whenever X increases, Y also increases, and when X decreases, Y decreases. Both the $r_{\text{States}} = .70$ and $r_{\Delta\text{Continuous}} = .98$ metrics agree that

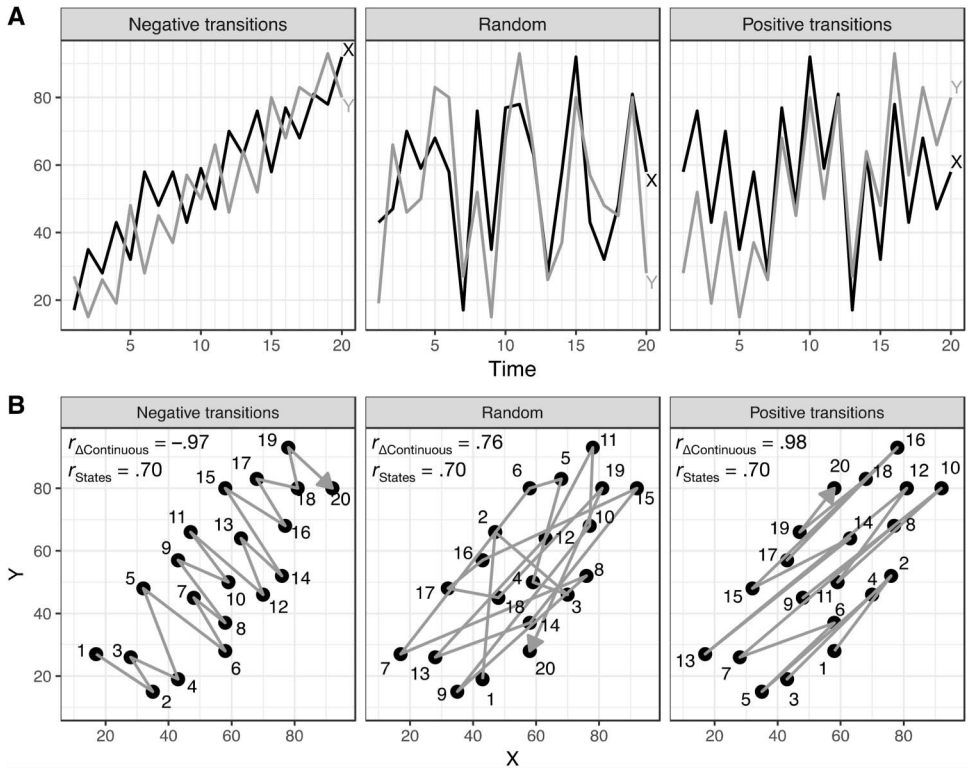

Figure 3. Sample dataset of a cause, X, and effect, Y, rearranged in three different orders. (A) Time series presentation of the data showing X and Y over time. (B) Scatterplots of the dataset with numbers indicating the order of observations. For all three orders, r_{States} = .70 but $r_{\Delta Continuous}$ takes on different values depending on the order of the observations.

there is a positive relation. In the random transitions panels in Figure 3, the same 20 observations were randomly ordered. In this condition, r_{States} = .70 and $r_{\Delta Continuous}$ = .73.

Experiments 1A and 1B tested whether people use transitions (difference scores) to control for temporal trends in time series data by comparing their judgments of causal strength in data sets like those in Figure 3, in which the states were held constant but the transitions were varied. If people only use the states (raw scores) for making judgments, then they would give similar causal strength judgments for all three data sets. If they use transitions, then they would give the highest judgments for the positive transitions condition, the lowest judgments for the negative transitions condition, and the random condition would be in the middle, though closer to the positive transitions condition.

The 20 observations in each dataset were presented sequentially in a standard trial-by-trial learning format. In Experiment 1A, the new scores for X and Y were displayed simultaneously on each trial. In Experiment 1B, there was a short delay between the display of the new score for X and the new score for Y on each trial. The reason for introducing the delay in Experiment 1B is that in many real-world situations there is some degree of a delay between the change in a cause and the change in an effect. We wanted to demonstrate that the proposed theory of learning from transitions can also be implemented in the context of a time series with delays.[3]

Method

Participants. For each experiment, 50 unique participants were recruited using Amazon Mechanical Turk (MTurk) and paid $0.60. The experiment lasted approximately 5 min. One additional participant completed Experiment 1B but did not claim payment. We included data from this participant.

Design and stimuli. The design and stimuli used in both experiments were identical. Participants inferred causal strength from data sets consisting of 20 observations of X and Y, in which X and Y could take on values ranging from 0 to 100. We manipulated the correlation between the states of X and Y as well as the transitions in a 2 (positive vs. negative r_{States}) × 3 (negative vs. random vs. positive transitions) within-subjects design.

Twenty data sets with r_{States} = .70 (e.g., Figure 3) were generated for the positive r_{States} condition using the *corgen* function from the R package *ecodist*. Many of the data sets were slightly modified in order to produce the desired patterns of transitions described below; however, the r_{States} value was always very close

[3] Though it would be theoretically possible to implement this strategy with long delays, for example, if the change of X at time *t* causes a change in Y at time *t* + 3, we expect that longer delays will become harder to parse in trial-by-trial paradigms. The point of Experiment 1B is simply that learning from transitions is theoretically possible even when there are delays.

to .70. Next, copies of each dataset were made by flipping the values of X around the midpoint of the scale (X = 50), creating data sets with r_{States} = -.70 for the negative r_{States} condition.

The observations from each of the generated data sets were reordered to produce three conditions (as in Figure 3). In the negative transitions condition, the trials were reordered by hand such that increases in X were always accompanied by decreases in Y, and vice versa. This process was accomplished roughly in the following way. Each dataset was plotted in scatterplot form (Figure 3B), and a diagonal line with a negative slope was drawn in the lower left-hand quadrant. The line was moved toward the right, and the order in which the observations intersected with the line, roughly, determined the order in the dataset. For this reason, all the transitions had a fairly similar, negative slope in the negative transitions condition.

In the positive transitions condition, the trials were reordered by hand such that increases in X were always accompanied by increases in Y. This was accomplished in a similar way as the negative transitions condition, except that the line that was drawn had a positive slope and was moved toward the left. This meant that all the transitions in the positive transitions condition had a fairly similar, positive slope.

In the random transitions condition, the order of the 20 observations was randomized by computer once; the same randomized order was viewed by every participant the dataset was presented to. This randomization resulted in a mix of positive and negative transitions; however, due to the states, most of the transitions in the negative r_{States} condition were negative, and most of the transitions in the positive r_{States} condition were positive. The correlation of difference scores ($r_{\Delta Continuous}$) was used as a measure of how positive or negative the transitions were (see Table 1).

Another way to think about the temporal trends present in the data sets is through the autocorrelation of X and Y. When the states and transitions are in conflict (e.g., the condition with r_{States} = .70 and negative transitions in Figure 3), X and Y both increase over time. This increase means that there is strong positive first-order autocorrelation (see Table 2). In the random transitions conditions, the autocorrelations are slightly negative. (We confirmed through simulation that the sampling distribution of the autocorrelation function of 20 randomly generated observations is in fact slightly negative.) In contrast, for most of the other conditions, there are no considerable increasing or decreasing trends, so the autocorrelation is roughly near zero on average. When the states and transitions are in agreement (e.g., the condition with r_{States} = .70 and positive transitions in Figure 3), the X variable actually has a fairly strong negative autocorrelation—this negative autocorrelation arises from the requirement to have X and Y change in the same direction. To be clear, even though the autocorrelations of the variables change across the conditions, this is not a confound—it is inherent to the fact that the conditions have different temporal trends. Strong temporal trends inherently have strong autocorrelation.

Table 1
Means (Standard Deviations) of $r_{\Delta Continuous}$ for Stimuli in Experiments 1 and 2

Transitions	*M* (*SD*) of $r_{\Delta Continuous}$	
	r_{States} = .70	r_{States} = −.70
Positive	.97 (.01)	.96 (.02)
Random	.73 (.07)	−.73 (.07)
Negative	−.96 (.02)	−.97 (.01)

Table 2
Means (Standard Deviations) of Autocorrelations for Stimuli in the r_{States} = .70 Conditions in Experiments 1 and 2

Transitions	*M* (*SD*) of autocorrelation when r_{States} = .70	
	X	Y
Positive	−.61 (.09)	−.004 (.27)
Random	−.15 (.13)	−.12 (.26)
Negative	.70 (.03)	.71 (.04)

Note. Autocorrelations are at a time lag of one trial. When r_{States} = −.70, the autocorrelations for the negative and positive transitions conditions are swapped; the autocorrelation is dependent on whether the states and transitions have the same polarity.

In Figure 3B, for the negative and positive transitions conditions, the observations start in the bottom and move to the top of the scatterplot; this is especially true for conditions in which the states and transitions conflict. We had intended to randomly present the data in either this order (1–20), or in the reverse order (20–1). However, due to a programming error, all participants saw the forward direction. We do not think that this is a critical issue because, if the order mattered at all, it is most plausible that the strong increasing trends in the negative transitions and positive states condition could lead participants to judge that there is a positive causal relation, which works against our hypothesis that they will judge a negative relation due to the negative transitions. This issue was fixed for subsequent experiments.

Each participant viewed one randomly chosen dataset (out of 20) within each of the six conditions, and the order of the conditions was randomized.

Procedure. Participants were told they would evaluate how the dosage of a drug (X) affected the size of a microorganism (Y) over 20 observations ("days"). Each condition was presented as a different drug-microorganism pair. On each day, a new dosage of the drug was administered to the microorganism and the size of the microorganism was observed under a microscope. The microorganism was represented using a circle, and there was also a triangle representing the needle used to inject the drug (see Figure 4). The dosage of the drug (X) was mapped onto the opacities of the microorganism and the needle. Darker shades represented higher doses of the drug; when X = 0, the stimuli were white (0% opacity), and when X = 100, the stimuli were black (100% opacity). The size of the microorganism (Y) was mapped to the diameter of the circle. The minimum diameter was set to be 30 pixels (approximately 0.79 cm, although this could vary slightly depending on participants' browser dimensions) and the maximum diameter was set to be 210 pixels (approximately 5.56 cm). Values of Y linearly determined the circle's diameter within that range (e.g., if Y = 50, the diameter was halfway between the minimum and maximum).

After each observation, there was a one second delay before a button appeared. When participants clicked the button, the next observation was shown. In Experiment 1A, the dosage of the drug

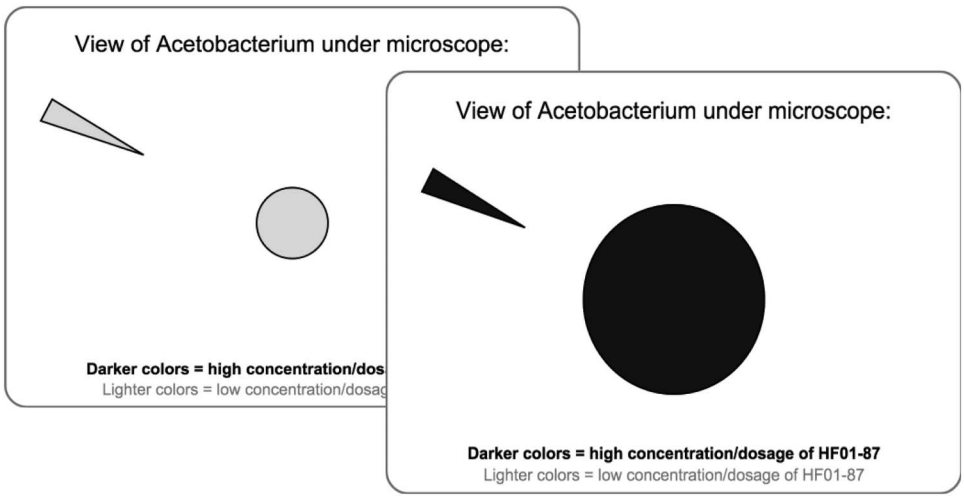

Figure 4. Presentation of stimuli in the visual format used in Experiments 1 and 4. Two observations are displayed to show how a transition might appear.

(X) and size (Y) of the microorganism changed simultaneously when the button was clicked. In Experiment 1B, after the button was clicked, the opacity of the needle changed, followed by a 250-ms delay after which the opacity of the microorganism changed, followed by a 750-ms delay after which the size of the microorganism changed; this sequence represented the causal sequence of the events in the cover story.

After 20 observations, participants judged the causal strength, on a scale from 8 (*high levels of the drug strongly cause the microorganism to increase in size*) to −8 (*high levels of the drug strongly cause the microorganism to decrease in size*). A rating of zero indicated no causal relationship.

Results

Effects of states and transitions. We tested whether participants used states and or transitions for estimating causal strength by testing whether r_{States} and or $r_{\Delta\text{Continuous}}$ were significant predictors using regression. The values of $r_{\Delta\text{Continuous}}$ corresponded to the particular dataset that a participant viewed. All the data sets in the positive and negative transitions conditions had $r_{\Delta\text{Continuous}}$ values of almost exactly .97 or −.96; however, there was more variance in the random condition (see Table 1).

We performed two sets of regressions. The first set of regressions tested the bivariate fits between each model and participants' judgments. The second set were multivariate regressions testing the influence of each model controlling for the other. The regressions had a by-participant random intercept for repeated measures, and by-participant random slopes for each predictor present in the given model to capture the possibility that some participants' judgments might be better predicted by r_{States} or $r_{\Delta\text{Continuous}}$.

The results of the regressions are reported in Table 3. In both Experiments 1A and 1B, and in both the bivariate and multivariate analyses, both r_{States} and $r_{\Delta\text{Continuous}}$ significantly predicted participants' causal strength judgments. Table 3 also reports r^2 for the bivariate analyses and partial-R^2 for multivariate analyses (in addition, we report effects sizes for all analyses in d). Transitions ($r_{\Delta\text{Continuous}}$) always accounted for more variance in participants' causal strength judgments than states (r_{States}).

The mean causal strength judgments for each condition in both Experiments 1A and 1B are displayed in Figure 5. Across both

Table 3
Model Fits for r_{States} and $r_{\Delta Continuous}$ in Regressions of Experiments 1A, 1B, and 2

		Predictor							
		r_{States}				$r_{\Delta\text{Continuous}}$			
Model	Experiment	*B* (*SE*)	*p*	r^2	*d*	*B* (*SE*)	*p*	r^2	*d*
Bivariate	1A	2.28 (.41)	<.001	.10	.67	2.80 (.42)	<.001	.24	1.12
	1B	2.53 (.41)	<.001	.11	.70	3.68 (.34)	<.001	.38	1.57
	2 (visual)	3.16 (.34)	<.001	.21	1.03	3.19 (.31)	<.001	.35	1.47
	2 (numerical)	2.78 (.37)	<.001	.21	1.03	1.56 (.27)	<.001	.11	.70
Multivariate: $r_{\text{States}} + r_{\Delta\text{Continuous}}$	1A	1.39 (.33)	<.001	.04	.41	2.49 (.42)	<.001	.19	.97
	1B	1.32 (.32)	<.001	.04	.41	3.39 (.33)	<.001	.33	1.40
	2 (visual)	2.19 (.37)	<.001	.15	.84	2.71 (.35)	<.001	.30	1.31
	2 (numerical)	2.42 (.37)	<.001	.16	.87	1.03 (.26)	<.001	.05	.46

Note. In multivariate models, r^2 represents partial-R^2.

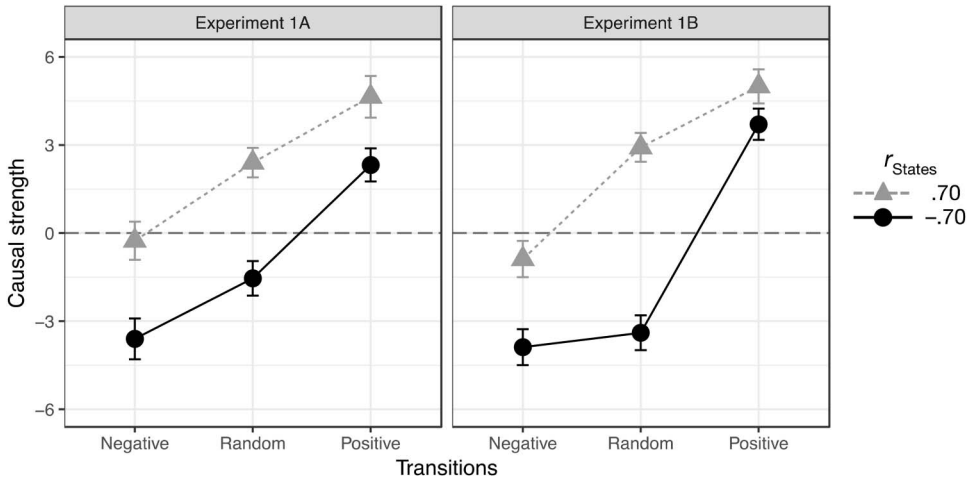

Figure 5. Condition means for Experiments 1A and 1B. Error bars represent standard errors.

experiments, the effect of states can be seen in the higher judgments for the $r_{States} = .70$ than $-.70$ conditions, and the effect of transitions can be seen in the increasing judgments from the negative to positive transitions conditions.

In Experiment 1A, the average of participants' judgments was above zero in the condition with negative r_{States} but positive transitions, $t(49) = 4.11$, $p < .001$; the effect of transitions "overrode" the effect of states. This same pattern was found in Experiment 1B; $t(50) = 6.97$, $p < .001$. In conditions with positive r_{States} but negative transitions, the average of participants' judgments was not significantly different from zero ($p = .69$ in Experiment 1A and $p = .08$ in Experiment 1B); the negative transitions "neutralized" the effect of the positive states.

Participant-level use of strategies. Although there was an overall effect of transitions in both experiments, we wanted to test if there were individual differences—perhaps only a minority of participants made judgments influenced by transitions while others focused on states. Across both experiments, we computed each participant's *transition score*. For each participant, we computed their mean judgment for the two conditions with positive transitions and for the two conditions with negative transitions (omitting their judgments in the two conditions with random transitions). The transition score was the difference between these two means. If participants distinguished between positive versus negative transitions as predicted, their transition scores would be positive (higher judgments in conditions with positive transitions). We also computed each participant's *state score*: We computed their mean judgment for the two conditions with positive states and the two conditions with negative states (omitting their judgments from the two conditions with random transitions). The state score was the difference between these two means. Participants who distinguished between positive versus negative states would have positive state scores.

The top row of Figure 6 plots the distribution of each participant's transition and state scores for Experiments 1A and 1B. Points in the upper half of each plot represent participants who exhibited a positive effect of transitions (transition scores >0). Participants in the right half of each plot represent participants who exhibited a positive effect of states (state scores >0). The number of participants falling in each quadrant is displayed in each panel of Figure 6.

Across both experiments, most participants (38 of 50 in Experiment 1A and 45 of 51 participants in Experiment 1B) exhibited a positive effect of transitions. A binomial test indicated that these proportions were significantly greater than chance (p's $< .001$). The effect of states was also positive in most participants: 38 of 50 in Experiment 1A ($p < .001$) and 37 of 51 in Experiment 1B ($p = .002$). In Experiment 1A, there was no significant correlation between participants' transition and state scores, $r = .18$, $p = .22$. In Experiment 1B, there was a marginal correlation, $r = .29$, $p = .04$. In sum, the effect of transitions was not confined to a small

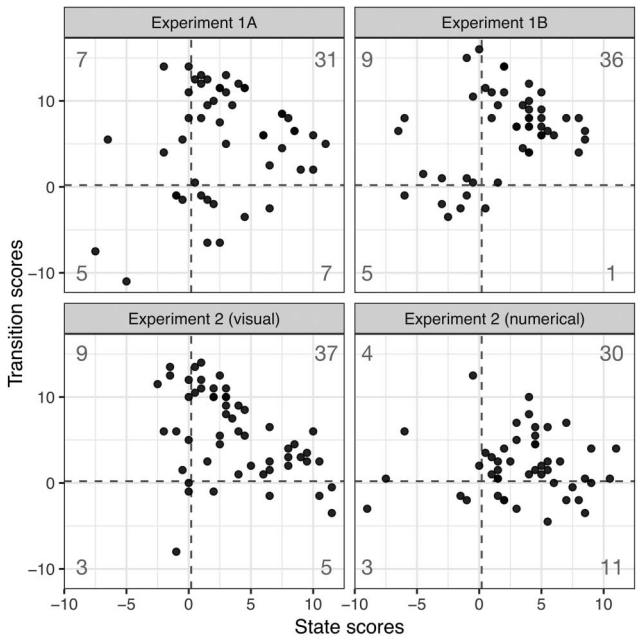

Figure 6. Transition and state scores for participants in Experiments 1A, 1B, and 2.

proportion of participants but was found in a majority. Furthermore, most participants exhibited both effects, and if anything, there is a positive rather than negative correlation between using the two strategies.

Discussion

Experiments 1A and 1B demonstrated that participants used transitions more than states for inferring causal strength from time series data with a continuous cause and continuous effect. Focusing on transitions helped participants uncover the true causal strength in this nonstationary time series setting. Though not tested directly because they are from separate experiments, the effect of transitions, if anything, was greater in Experiment 1B than 1A. It is possible that the delay highlighted the time series nature of the task and led to more use of transitions.

Experiment 2: Symbolic Versus Perceptual Stimuli Formats

In Experiment 2, we tested whether the use of transitions for inferring causal strength is moderated by the perceptual nature of the stimuli. We hypothesized that naturalistic perceptual presentations such as a visual format may lead to a greater focus on changes than a symbolic numerical format for the reasons described in the introduction.

Method

Participants. One-hundred and two participants were recruited using Amazon MTurk and Paid $0.60. The experiment lasted about 5 min.

Design. The same design and data sets as Experiment 1A was used, except that an additional between-subjects factor was added so that half the participants viewed the data presented in a visual format, while half the participants viewed the data in a numerical format.

Stimuli. The stimuli was largely similar to the prior experiments, except for the presentation formats (see Figure 7). In the numerical condition, the stimuli were presented as numbers on a scale 0–100. In the visual format, both the cause and effect were displayed using vertical sliders (see Figure 7A).[4] The sliders had a height of 280 pixels (approximately 7.41 cm).

Additionally, the order of trials was counterbalanced to move either forward or in reverse. For example, some participants saw the order 1–20 in Figure 3, and others saw the order 20–1. This was determined randomly for each scenario a participant saw.

Results

Effects of states and transitions. There were no differences in participants' judgments between data sets that were presented in a forward versus reversed order, so we analyzed all data sets together. Means for all conditions are presented in Figure 8. We first ran the same sets of analyses as in Experiment 1, separately for the two conditions (see Table 3). In the bivariate analysis, both r_{States} and $r_{\Delta Continuous}$ significantly predicted participants' causal strength judgments. The variance explained was larger for $r_{\Delta Continuous}$ in the visual condition, and larger for r_{States} in the numerical condition. The effect of r_{States} remains the same across conditions, but the effect of $r_{\Delta Continuous}$ is much stronger in the visual than numerical condition. The same pattern held up for the multivariate analysis; both predictors were significant in both conditions, and the variance explained by the two predictors showed the same switch. The larger effect of transitions in the visual condition can be seen in Figure 8 in the steeper slopes of the lines from the negative to positive condition in the visual compared with numerical condition.

To formally test whether there was an interaction between the two predictors and the presentation format, we ran a regression with five predictors: r_{States}, $r_{\Delta Continuous}$, presentation format, the interaction between r_{States} and presentation format, and the interaction between $r_{\Delta Continuous}$ and presentation format. Due to repeated measures, there was a by-participant random intercept and a random slope for r_{States} and $r_{\Delta Continuous}$ (the within-subjects predictors). In this regression, there was a significant overall effect of r_{States} ($B = 2.19$, $SE = 0.36$, $p < .001$, partial-$R^2 = .15$, $d = 0.84$) and a significant overall effect of $r_{\Delta Continuous}$ ($B = 2.71$, $SE = 0.30$, $p < .001$, partial-$R^2 = .17$, $d = 0.91$). There was no significant interaction between r_{States} and presentation format ($B = 0.22$, $SE = 0.53$, $p = .67$, partial-$R^2 = .001$, $d = 0.06$). Most importantly for this study, there was a significant interaction between $r_{\Delta Continuous}$ and presentation format ($B = -1.68$, $SE = 0.44$, $p = .001$, partial-$R^2 = .04$, $d = 0.41$); the negative coefficient means that the effect of transitions was larger in the visual condition.

Participant-level use of strategies. Figure 6 (bottom row) displays participants' transition and state scores, calculated the same way as in Experiment 1. Similar to Experiments 1A and 1B, the effect of transitions was not confined to a small proportion of participants. In each condition, the proportion of participants exhibiting a positive effect of transitions was above chance: 46 of 54 in the visual condition ($p < .001$), and 34 of 48 in the numerical condition ($p = .006$). A chi-square test of independence found these proportions did not differ across the two conditions ($p = .13$). A positive effect of states was also exhibited by a majority of participants: 42 of 54 in the visual condition, and 41 of 48 in the numerical condition (p's $< .001$). These proportions did not differ across the two conditions ($p = .46$). There was a correlation between participants' state and transition scores in the visual condition, $r = -.46$, $p < .001$ but not the numerical condition, $r = .01$, $p = .96$. The correlation between participants' state and transition scores in the visual condition suggest that participants exhibited either a strong effect of transitions or states, but not both simultaneously.

Discussion

In sum, participants used transitions for estimating causal strength more in the visual than in the numerical condition. This meant that their judgments were more accurate when the stimuli were presented visually rather than numerically. We use the word accurate because in the negative (positive) transitions condition, the true causal strength is negative (positive) once the temporal confound is accounted for, and

[4] We used linear sliders instead of the opacity and circle in Experiment 1, out of concern that these might not be viewed linearly; opacity might not be perceived linearly, and the circle's size could be interpreted in terms of either diameter or area. We thank two reviewers for pointing out these potential issues.

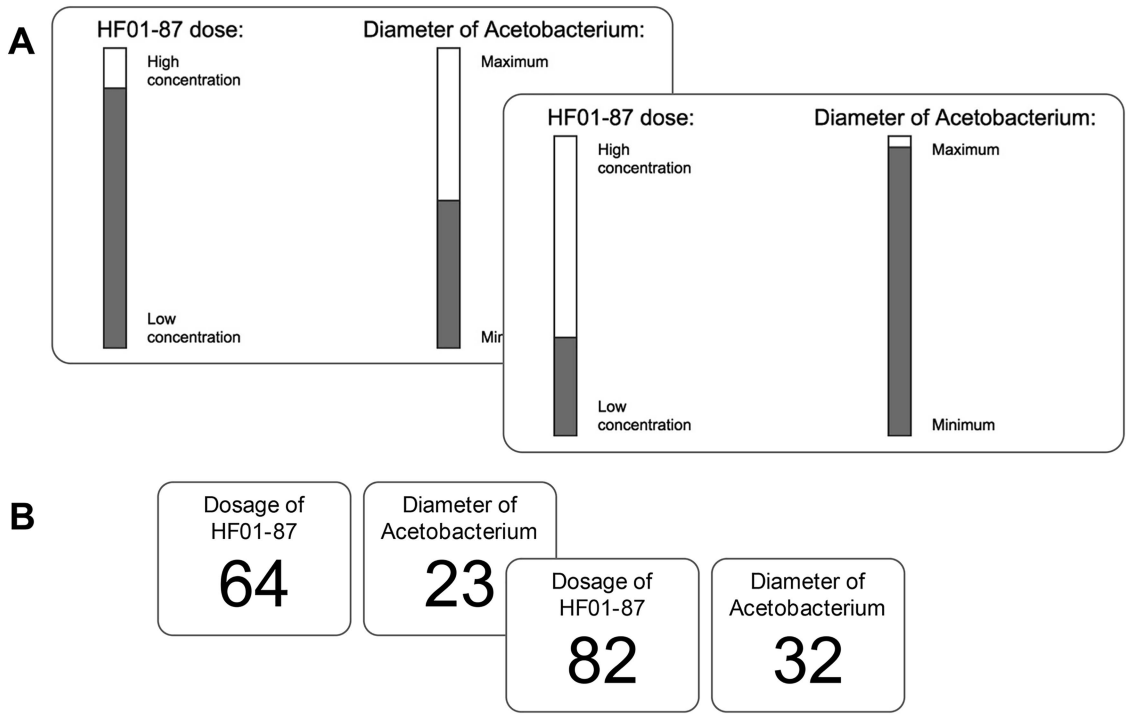

Figure 7. (A) Presentation of stimuli in the visual format used in Experiments 2 and 3. (B) Presentation of stimuli in the numerical format used in Experiment 2. Two observations are displayed to show how a transition might appear.

participants' judgments more closely tracked the difference between negative versus positive transitions in the visual condition.

This finding is somewhat paradoxical: humans are able to represent numbers precisely, an ability that enables finer discrimination between quantities and complex computations, but this ability appears to interfere with causal learning in time series contexts. In contrast, humans appear to be fairly well adapted to infer causality correctly in time series contexts from naturalistic stimuli. Collectively, the findings from Experiments 1A, 1B, and 2 demonstrate a robust effect of transitions (even across two different visual presentation formats and a numerical format). Although the effect of transitions can be moderated by perceptual factors, the effect was present in all conditions.

Experiment 3: Ruling out Primacy and Recency Effects as an Explanation for the Effect of Transitions

A potential alternative explanation for the results of the prior experiments is that participants' causal inferences are influenced by primacy or recency effects (Collins & Shanks, 2002; Dennis & Ahn, 2001; Fernbach & Sloman, 2009; Glautier, 2008).[5] A primacy effect means that participants' judgments are based primarily on the first few trials. This could happen if participants quickly form beliefs about causal strength and then discount or ignore later evidence, or interpret it in a way that is consistent with their initial beliefs. A recency effect means that participants' judgments are based primarily on the last few trials. Recency effects can happen due to limited memory, or reflect a rational attempt to update an estimate for the most recent context in a nonstationary setting.

In Experiments 1 and 2, both recency and primacy effects were confounded with $r_{\Delta\text{Continuous}}$. For example, consider the data in Figure 3. Even though $r_{\text{States}} = .70$ for the entire dataset, in the negative transitions condition, $r_{\text{States}} = -.84$ for the first three observations, and $r_{\text{States}} = -.67$ for the last three observations. (For any window of three observations, r_{States} is very negative, but with larger windows, becomes more positive.) In the random transitions condition, r_{States} is usually positive for small windows, and in the positive transitions condition, r_{States} is always positive for small windows.

Experiment 3 was conducted to rule out the possibility that primacy or recency effects might explain the findings we attributed to transitions. To do this, we created stimuli in which the primacy and recency effects conflicted (e.g., r_{States} was positive for the first few trials and negative for the last few, or vice versa). $r_{\Delta\text{Continuous}}$ was either consistent with a primacy effect or a recency effect, but never both. This allowed us to test how participants' judgments aligned with $r_{\Delta\text{Continuous}}$ or with primacy/recency effects.

Method

Participants. One-hundred participants were recruited from Amazon MTurk and paid $1.50. This experiment lasted between 5 and 10 min. One additional participant completed the experiment but did not claim payment. We included data from this participant.

Design and stimuli. The study used a 2 (positive vs. negative transitions) × 2 (positive vs. negative recency) within-subjects

[5] We thank an anonymous reviewer for suggesting this explanation.

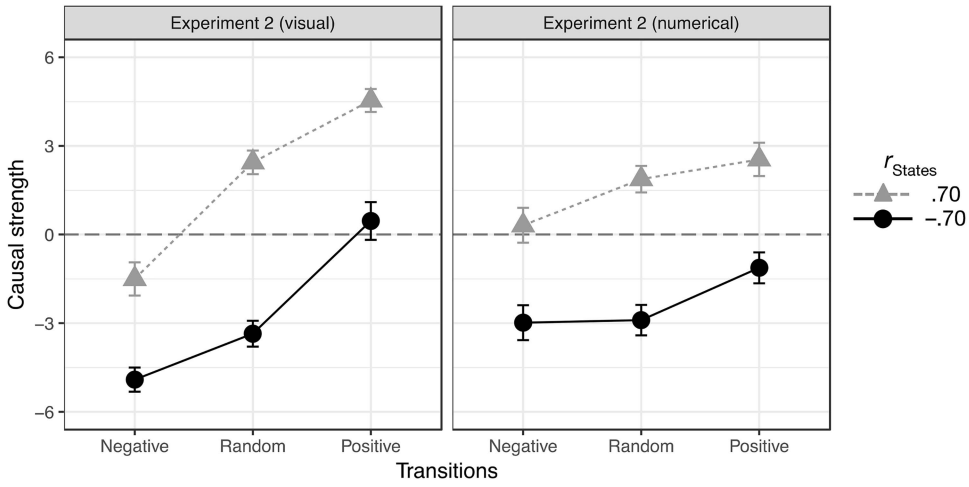

Figure 8. Condition means for Experiment 2. Error bars represent standard errors.

design. The recency factor was negatively related to primacy—positive recency conditions had negative primacy, and vice versa. This design was accomplished in the following way. We created data sets with 17 observations in which the states were held constant, but varied whether either a recency or primacy effect of the states was consistent with the transitions. Figure 9 shows an example of one of the data sets.

These data sets have a number of unique features. First, the transitions within each dataset were always positive or always negative (with one exception explained later). In Figure 9, the transitions are negative. Second, the data sets were symmetric, which meant that r_{States} for all 17 observations was zero. One variable underwent a positive or negative trend (X increased in Figure 9), and the other variable (Y in Figure 9) underwent a positive trend followed by a negative trend, or vice versa. Due to this second feature, the chosen order of observations (either from 1–17 or 17–1) ensured that r_{States} was negative in the first half and positive in the second half (as in Figure 9), or vice versa. The order of observations in Figure 9 also ensured that r_{States} calculated with a recency effect was always $\leq$ zero and r_{States} calculated with a primacy effect was always $\geq$ zero. We allowed for the possibility that a recency effect could be calculated by taking the r_{States} value of the last n observations. For example, r_{States} could be calculated for observations 16–17, or 15–17, . . . , or 2–17. In Figure 9, all the r_{States} values calculated with a recency effect are positive with two exceptions; for observations 16–17, the r_{States} value cannot be calculated because they have the same X value, and for observations 15–17, $r_{States} = 0$. We found these exceptions to be necessary to achieve all the stimuli features we needed. The same was done for a primacy effect, which could be calculated by taking the r_{States} value of the first n observations (e.g., 1–2, or 1–3, . . . , or 1–16). In Figure 9, all the r_{States} values calculated with a primacy effect are negative.

From the basic dataset in Figure 9, we made 16 versions by combining three ways of manipulating the data. First, we reversed the order of the 17 observations; doing so to Figure 9 meant that all transitions were still negative, but a recency effect would be negative and a primacy effect would be positive. Second, we flipped the X observations around the midpoint; doing so to the data in Figure 9 meant that the transitions were positive instead of negative. Lastly, we also swapped the values of X and Y, which meant that for half the data sets X had a linear trend and Y had both increasing and decreasing trends, whereas for the other half Y had a linear trend and X had both increasing and decreasing trends.

These 16 data sets are summarized in Table 4; the 2 × 2 design can be seen by focusing on the transitions and recency columns. In data sets with positive (negative) transitions, $r_{\Delta Continuous} = .81$ ($-.81$). Table 4 shows the possible ranges of r_{States} calculated with

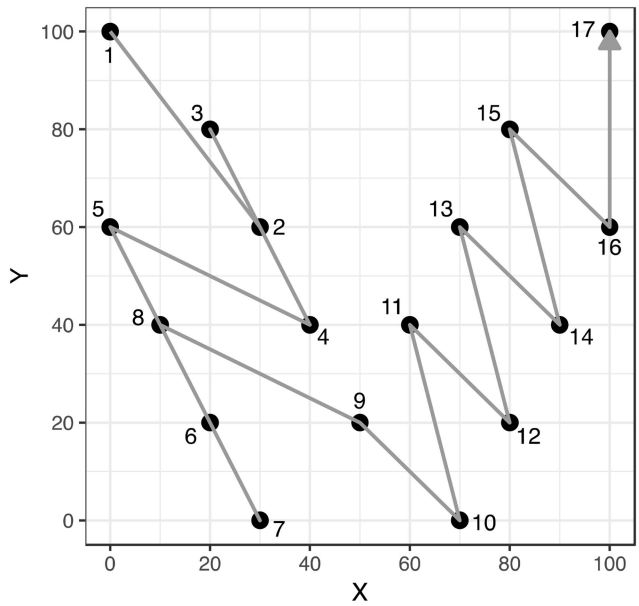

Figure 9. A sample dataset in Experiment 3. The numbers represent the order of the 17 observations. All the transitions are negative, except the last one which is neutral because X does not change. All calculations of a recency effect have $r_{States} \geq 0$, whereas all calculations of a primacy effect have $r_{States} < 0$. If the order of the trials is reversed, then all transitions would still be negative, but a recency (primacy) effect would be negative (positive).

Table 4
Stimuli Characteristics for Conditions in Experiment 3

Group	Dataset	Transitions	Trend X	Trend Y	Range of r_{States} calculated with Primacy	Range of r_{States} calculated with Recency
I	1	Negative	+	+ −	[0, .58]	[−1, −.20]
	2		−	− +		
	3		+ −	+		
	4		− +	−		
II	5	Negative	+	− +	[−1, −.20]	[0, .58]
	6		−	+ −		
	7		+ −	−		
	8		− +	+		
III	9	Positive	+	− +	[−.58, 0]	[.20, 1]
	10		−	+ −		
	11		+ −	−		
	12		− +	+		
IV	13	Positive	+	+ −	[.20, 1]	[−.58, 0]
	14		−	− +		
	15		+ −	+		
	16		− +	−		

a primacy and recency effect; all calculations of primacy and recency effects for all ranges of the observations have opposing signs. For all data sets, the first order autocorrelation of the variable with a linear trend was .78, and the autocorrelation of the variable that had both increasing and decreasing trends was .35.

Procedure. The procedure was identical to the visual condition in Experiment 2, except that participants experienced eight scenarios. For each participant, two data sets were randomly selected from each of the four groups in Table 4. The data sets were experienced in a random order.

Results

Transitions versus primacy/recency effects. Figure 10A displays the mean causal strength judgments by transitions and recency effects in the data sets. To test if judgments were influenced by transitions or by primacy/recency effects, we ran a regression with two predictors: transitions (either positive or negative) and recency effects in a data (either positive or negative). As a reminder, whenever the recency effects were positive, primacy effects were negative, and vice versa. The regression also included the interaction between these two predictors and a by-participant random intercept and random slopes for the predictors (due to repeated measures).

All three predictors, the slope for transitions, for recency, and the interaction, were entered into the model simultaneously. There was a significant effect of transitions; the positive transition condition had higher causal strength judgments than the negative transition condition ($B = 4.51$, $SE = 0.48$, $p < .001$, partial-$R^2 = .33$, $d = 1.40$). There was no main effect difference between stimuli with positive versus negative recency effects ($p = .55$, partial-$R^2 = .002$, $d = 0.09$). There was a small but significant interaction ($B = 1.08$, $SE = 0.43$, $p = .013$, partial-$R^2 = .006$, $d = 0.16$); the positive recency effect stimuli were judged to have a stronger causal strength in the positive transitions condition ($B = 0.90$, $SE = 0.35$, $p = .011$, partial-$R^2 = .01$, $d = 0.20$) but not the negative transitions condition ($p = .54$).

Participant-level analysis. Figure 10B displays each participant's transition score, calculated the same way as in the prior experiments. However, instead of a state score for each participant, we display each participant's *recency score.* We computed their mean judgment for conditions with positive recency effects and conditions with negative recency effects. The recency score was the difference between these two means. Participants who displayed recency effects would have positive recency scores. Conversely, participants who displayed primacy effects would have negative recency scores.

The majority of participants (86 of 101) displayed a positive effect of transitions ($p < .001$). The number of participants showing recency effects (51 of 101) was not greater than chance ($p = 1$). The majority of participants displayed small primacy/recency effects, but five participants displayed extreme recency effects with recency scores >5. There was a marginal negative relationship between participants' transition and recency scores, $r = -.19$, $p = .06$. Figure 10B suggests this relationship was due to a few participants displaying strong recency effects but no effect of transitions.

Discussion

The goal of Experiment 3 was to rule out primacy and recency effects of r_{States} as potential explanations for the results in Experi-

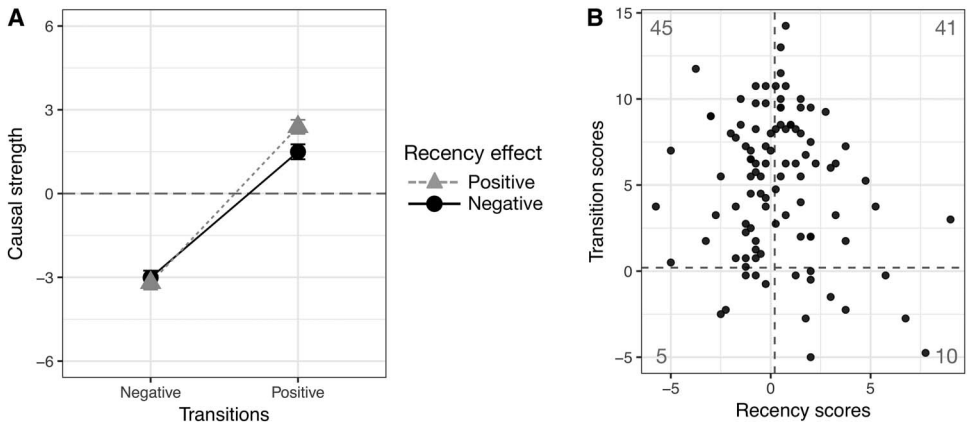

Figure 10. (A) Condition means for Experiment 3. Error bars represent standard errors. (B) Transition and recency scores for participants in Experiment 3.

ments 1 and 2. In Experiment 3, there was a large main effect of transitions but no main effect of primacy/recency. There was a marginal interaction between transitions and primacy/recency effects of r_{States}, which was confined to conditions with positive transitions. Furthermore, the participant-level analysis suggested that this effect was confined to a small number of participants. In sum, primacy/recency effects of r_{States} cannot account for the large effect of transitions. Of course, it is possible that there are primacy or recency effects in regards to participants' use of $r_{\Delta Continuous}$, though that is not something we study in the current research.

Experiments 4 and 5: Learning From the Magnitude or Direction of Transitions

In Experiments 4 and 5, we investigated whether transitions are encoded as magnitudes ($\Delta_{Continuous}$) or discrete values (Δ_{Binary}). As previously discussed, when a stimulus changes, the change can either be encoded on a continuous scale in terms of the amount of change, or on a binary scale of increase versus decrease. Encoding the change using a binary scale and inferring causal strength using a process similar to $r_{\Delta Binary}$ may serve as an easy-to-use heuristic that approximates $r_{\Delta Continuous}$ (see Appendix A, Part 2).

In Experiments 1 to 3, $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ were confounded, so it was impossible to distinguish them. Both models predicted very negative estimates of causal strength for the negative transitions conditions and very positive estimates for the positive transitions conditions. For Experiments 4 and 5, we used two different strategies to generate data sets in which $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ diverged, to estimate the unique effects of each model on participants' causal strength judgments above and beyond r_{States}. Unlike Experiments 1 to 3, none of the variables had linear temporal trends. From a normative perspective, this means that it is less necessary to account for time, so it is possible that the effects of transitions will be smaller. However, we found it necessary to move away from linear temporal trends in order to distinguish $r_{\Delta Continuous}$ and $r_{\Delta Binary}$.

In Experiment 4, this was accomplished using data sets from a type of Simpson's Paradox, resulting in stimuli with a low correlation between $r_{\Delta Binary}$ and $r_{\Delta Continuous}$, while holding r_{States} constant. The downside of this approach was that the data sets were somewhat unusual; however, the upside was that the three predictors could be differentiated well. In Experiment 5, we randomly generated data sets with a fixed value of r_{States}. Though $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ were positively correlated, their unique effects could be disentangled with regression.

To anticipate the results, both Experiments 4 and 5 found that participants used a strategy like $r_{\Delta Binary}$ for inferring causal strength, but there was no evidence that they used $r_{\Delta Continuous}$ once r_{States} was controlled for.

Experiment 4

Method

Participants. Fifty participants were recruited using Amazon MTurk and paid $1.50. The experiment spanned 12 scenarios and lasted about 10 min–12 min.

Stimuli generation. We created data sets that held r_{States} constant, but varied $r_{\Delta Binary}$ and $r_{\Delta Continuous}$. This was challenging because $r_{\Delta Binary}$ and $r_{\Delta Continuous}$ are typically highly correlated.

Each dataset had 12 observations—two observations of each of the following six states of (X, Y): (0, 20), (20, 0), (40, 60), (60, 40), (80, 100), and (100, 80). In creating these stimuli, we were careful to ensure that it was easy to perceptually distinguish all values on the X (opacity) and Y (size) dimensions. Using these 12 observations meant that for all stimuli, $r_{States} = .83$. The observations formed three clusters (I–III, see Figure 11A) with low, medium, or high values for X and Y. Transitions between two observations within a cluster (e.g., [0, 20] to [20, 0]) necessarily involved a negative transition (X increasing and Y decreasing, or vice versa). Transitions between clusters (e.g., [0, 20] to [60, 40]) necessarily involved a positive transition. This can be viewed as a type of Simpson's paradox; such data sets might occur if the cause and effect are negatively related once a third variable (the clusters) is held constant.

Within each dataset it was possible for $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ to diverge considerably depending on the observation order and the resulting transitions. Increasing the ratio of within to between-cluster transitions decreased both $r_{\Delta Continuous}$ and $r_{\Delta Binary}$. However, for a given ratio of within to between-cluster transitions, $r_{\Delta Continuous}$ was also influenced by the between-cluster transition path. Transition paths with smaller jumps between clusters produced lower $r_{\Delta Continuous}$ values. For example, in Dataset 1 of Figure 11, the between-cluster transitions are between adjacent clusters (Cluster I and II, or Cluster II and III), resulting in $r_{\Delta Continuous} = .19$. In contrast, transition paths with large jumps between clusters produced higher $r_{\Delta Continuous}$ values. For example, in Dataset 2 of Figure 11, there are three transitions between Clusters I and III, resulting in $r_{\Delta Continuous} = .65$.

Ten-thousand data sets were generated by randomly ordering the 12 observations, with the constraint that the same observation could not occur consecutively. The $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ values for these 10,000 data sets are plotted in Figure 12A as black circles. For each of these data sets, we created versions for which $r_{States} = -.83$ by flipping the values of X around the midpoint of 50. This procedure also flipped the $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ values of that dataset, which are plotted in Figure 12A as gray triangles.

To discriminate between $r_{\Delta Continuous}$ and $r_{\Delta Binary}$, we selected data sets from regions on the periphery of the distributions of original and flipped data sets, giving us 2,760 data sets in total. In Figure 12A, the original regions are marked with solid squares, and the corresponding flipped regions are marked with dashed squares. The numerical labels of the regions correspond roughly to the $r_{\Delta Binary}$ value of the data sets in the region. The high-low label ("H" vs. "L") refers to whether the $r_{\Delta Continuous}$ value is high or low within a given numerical region. Data sets in Region 1 correspond to data sets in Region 7 and so on, with swapped high-low labels. For example, data sets in Region 1H with $r_{States} = .83$ have corresponding reversed data sets with $r_{States} = -.83$ in Region 7L.

The data sets that were not selected tended to have many more between-cluster transitions overall. Only those with sufficient local within-cluster transitions in the opposite direction of the larger between-cluster transitions gave rise to Simpson's Paradox, yielding diverging values for $r_{\Delta Continuous}$ and $r_{\Delta Binary}$.

Each participant was presented with 12 data sets during the experiment—one from each of the original regions or their corresponding flipped regions. For a particular region, participants were randomly shown either the $r_{States} = .83$ dataset or the $r_{States} = -.83$ version (either from Region 1L or Region 7H, Region 2L or Region 8H, etc.).

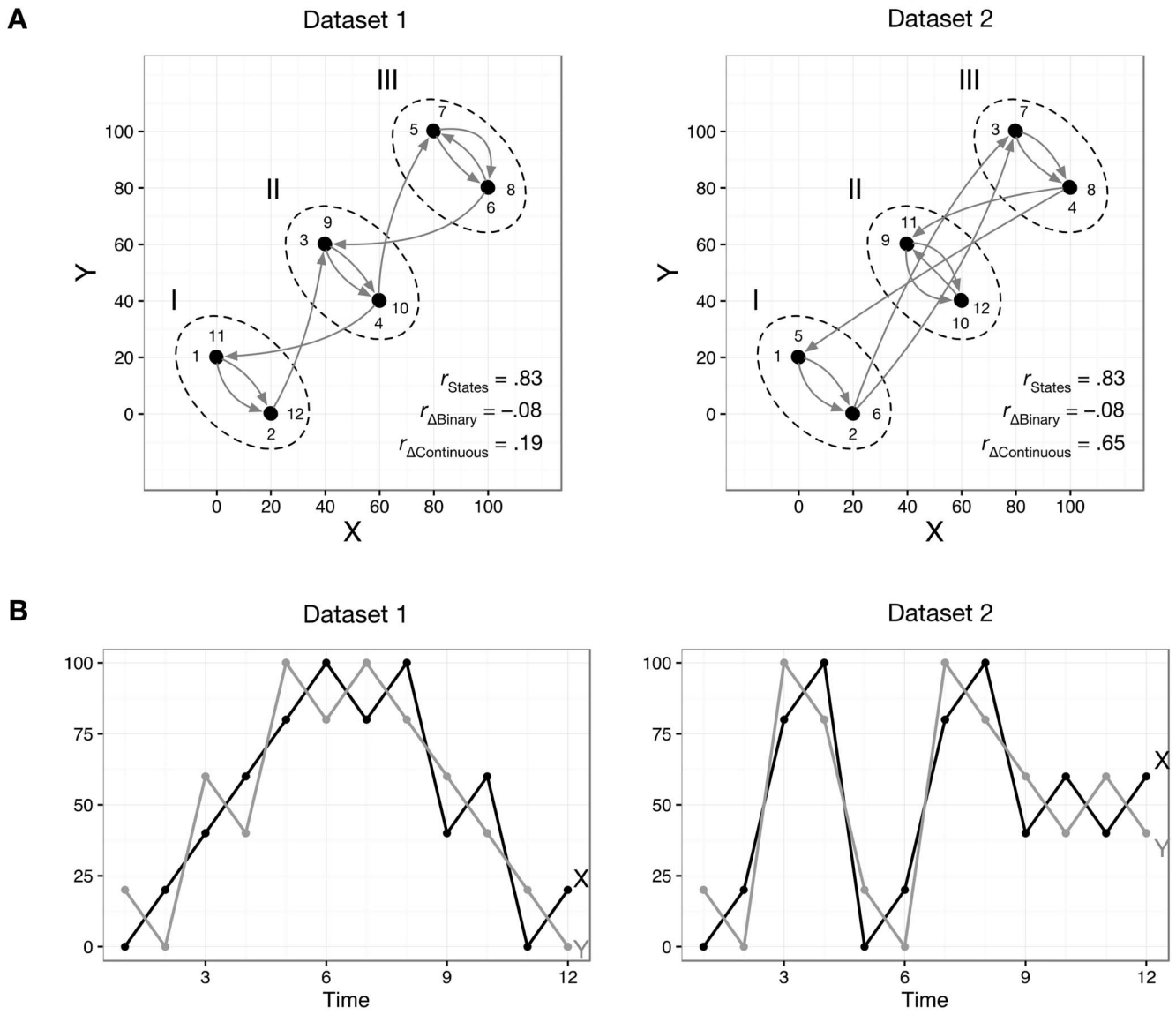

Figure 11. Two sample data sets from Experiment 4. Each scenario consisted of 12 observations (each data point was observed twice). (A) Scatterplots of the data sets. Numbers reflect the order of the observations. (B) Time series presentation of the data sets.

Stimuli properties. The average autocorrelations of X and Y in all data sets differed across regions (see Table 5). In the r_{States} = .83 data sets, the autocorrelations were lower for higher region numbers (higher $r_{\Delta Continuous}$ and $r_{\Delta Binary}$) due to the increased number of larger between-cluster transitions. For the same reason, the autocorrelations were also lower for the "H" regions (with higher $r_{\Delta Continuous}$) than for "L" regions. We view these differences in the autocorrelations not as a confound, but as another way to understand the differences between regions, which are useful for disentangling $r_{\Delta Binary}$ from $r_{\Delta Continuous}$.

The correlations between r_{States}, $r_{\Delta Continuous}$, and $r_{\Delta Binary}$ in the stimuli viewed by participants are shown in Table 6. The "Raw" row provides the correlations of the three predictors coming from all 24 regions in Figure 12A, which included data sets in which r_{States} = .83 and −.83. The correlation between r_{States} and $r_{\Delta Continuous}$ was relatively high because $r_{\Delta Continuous}$ is a good estimator of r_{States}. Crucially for the purposes of this experiment, the correlation between $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ was close to zero. The correlation between r_{States} and $r_{\Delta Binary}$ was actually negative, and can be seen in Figure 12A: Out of the 24 regions, the region with the highest $r_{\Delta Binary}$ value is in the r_{States} = −.83 condition (Region 7), and the region with the lowest $r_{\Delta Binary}$ value is in the r_{States} = .83 condition (Region 1). One implication of the negative correlation between r_{States} and $r_{\Delta Binary}$ was that the bivariate correlation between the $r_{\Delta Binary}$ values and participants' causal strength judgments was likely to be weak or even negative since r_{States} has a positive influence on causal strength judgments.

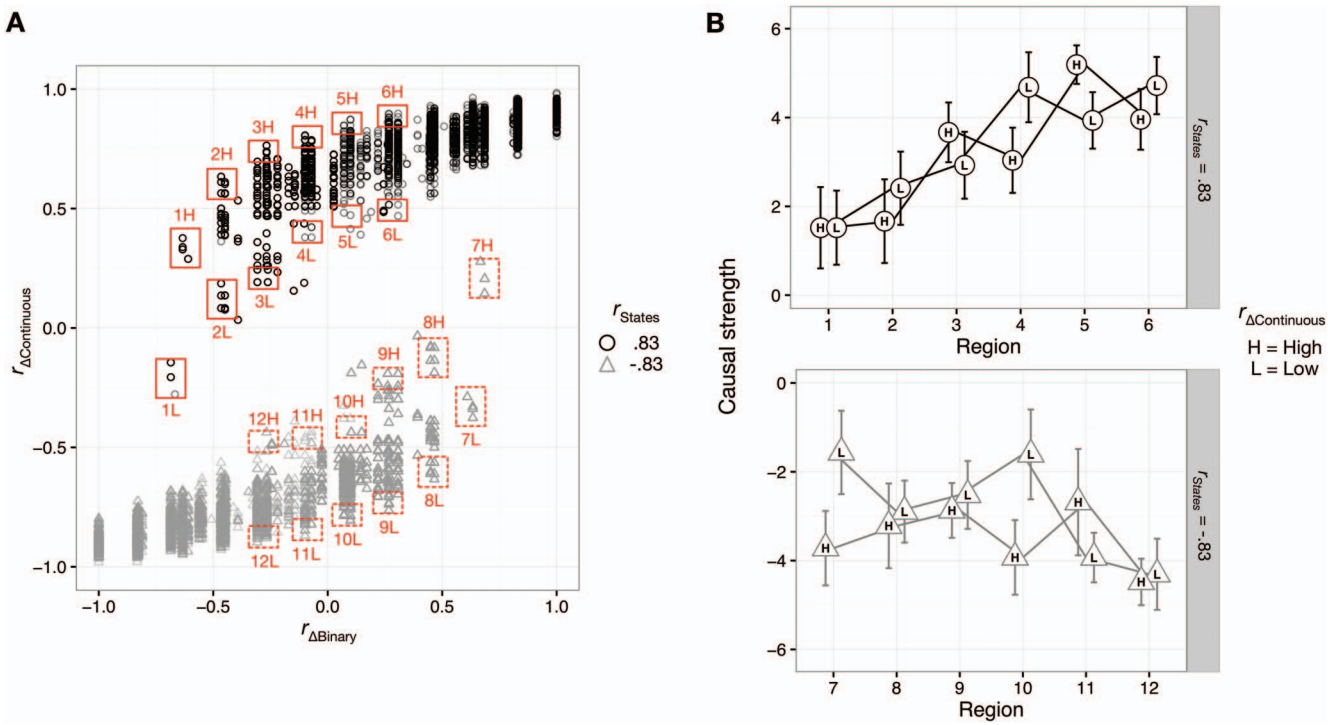

Figure 12. (A) $r_{\Delta Binary}$ and $r_{\Delta Continuous}$ values for 10,000 data sets with r_{States} = .83 and their reversed counterparts with r_{States} = −.83 generated for Experiment 4. Stimuli were sampled from the marked regions. (B) Mean causal strength judgments for stimuli from each region. Error bars reflect standard errors. See the online article for the color version of this figure.

Because r_{States} and $r_{\Delta Continuous}$ were fairly highly correlated, we analyzed the data in two ways. The first way used the raw scores of the three predictors. The second way involved recoding the r_{States} = −.83 stimuli back into the positive domain (described in detail below). This meant that we treated all data as if r_{States} = .83, which allowed us to estimate the effects of $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ without having to statistically control for r_{States}. The correlation between the $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ predictors among all data sets with r_{States} = .83 was .61 (see Table 6). We used regression to separately estimate the effects of these two correlated predictors. There are no correlations with r_{States} in the "Reverse" coding rows in Table 6 because once the data were recoded, r_{States} was fixed.

Procedure. The stimuli were presented using the visual format of Figure 4; the dosage of the drug (X) was mapped onto the opacity of the microorganism and the needle, and the size of the microorganism (Y) was mapped onto the diameter of the circle. The procedure was the same as in Experiment 1A, except that participants experienced 12 scenarios.

Results

Figure 12B plots the means of participants' causal strength judgments for scenarios from each of the 24 regions in Figure 12A. In the r_{States} = .83 condition, higher region numbers contained stimuli with higher $r_{\Delta Binary}$ values, and were associated with higher causal strength judgments (upward trending lines). However, the plot does not reveal a consistent tendency for participants to provide higher causal strength judgments for the $r_{\Delta Continuous}$ high versus low regions.

Table 5
Means (Standard Deviations) of Autocorrelations for Stimuli by Region in Experiment 4

Region	r_{States} = .83	1H	2H	3H	4H	5H	6H
	r_{States} = −.83	7L	8L	9L	10L	11L	12L
M (*SD*) of auto-correlation	X	.58 (.01)	.31 (.07)	.12 (.09)	.11 (.10)	−.16 (.13)	−.14 (.19)
	Y	.53 (.03)	.31 (.07)	.13 (.10)	.09 (.08)	−.14 (.19)	−.34 (.10)
Region	r_{States} = .83	1L	2L	3L	4L	5L	6L
	r_{States} = −.83	7H	8H	9H	10H	11H	12H
M (*SD*) of auto-correlation	X	.70 (.07)	.67 (.09)	.59 (.06)	.57 (.09)	.43 (.05)	.41 (.05)
	Y	.76 (.07)	.67 (.09)	.59 (.06)	.59 (.09)	.47 (.06)	.42 (.05)

Table 6
Correlations Between Predictors in Stimuli Viewed by Participants in Experiments 4 and 5

Experiment	Stimuli coding	Range of r_{States}	Correlations between predictors		
			$r_{States} \sim r_{\Delta Continuous}$	$r_{States} \sim r_{\Delta Binary}$	$r_{\Delta Continuous} \sim r_{\Delta Binary}$
4	Raw	.83 or −.83	.80	−.49	−.08
	Reverse	.83	—	—	.61
5	Raw	.50 or −.50	.94	.79	.85
	Reverse	.50	—	—	.52

Note. Raw-coding means that r_{States} values were both positive and negative. Reverse-coding means that the values on the X-axis were flipped around the midpoint, so all r_{States} values were positive.

In the $r_{States} = -.83$ condition, higher region numbers contained stimuli with lower $r_{\Delta Binary}$ values. Though the trend was not as clear as in the $r_{States} = .83$ condition, the average causal strength judgment seemed to decrease slightly with higher region numbers, which was the expected effect if $r_{\Delta Binary}$ had a positive effect on causal strength judgments. If anything, the effect of $r_{\Delta Continuous}$ appeared to go in the opposite direction: In Regions 7 and 10 higher $r_{\Delta Continuous}$ values had lower average causal strength judgments.

In the following sections, we analyze these patterns formally with inferential statistics in two different ways. In the first set of analyses, we statistically controlled for r_{States}, and in the second set of analyses we held r_{States} constant by reverse-coding the $r_{States} = -.83$ stimuli.

Raw-coding analysis. The r_{States}, $r_{\Delta Continuous}$, and $r_{\Delta Binary}$ model predictions for each dataset were used as predictors of causal strength judgments in a series of regression models. The first set of models were bivariate regressions between each predictor and participants' judgments. The second set of models estimated the effect of $r_{\Delta Continuous}$ controlling for r_{States}, and the effect of $r_{\Delta Binary}$ controlling for r_{States}. The third set of multivariate regressions included all three predictors. The regressions had a by-participant random intercept for repeated measures, and by-participant random slopes for each predictor present in the given regression to capture the possibility that some participants' judgments might be better predicted by a particular model. The results are reported in Table 7. A bivariate analysis found that all three predictors were significant. However, the coefficient for $r_{\Delta Binary}$ was negative. This effect was anticipated in the analysis of the stimuli properties; it is due to the fact that the r_{States} values are negatively correlated with the $r_{\Delta Binary}$ values in the stimuli.

In the second set of regressions that controlled for r_{States}, $r_{\Delta Binary}$ was highly significant, and $r_{\Delta Continuous}$ was marginally significant. In the full multivariate regression with all three predictors, $r_{\Delta Binary}$ was highly significant, and $r_{\Delta Continuous}$ was not significant.

The effect of $r_{\Delta Binary}$ on causal strength judgments can be seen in Figure 13, a scatterplot of $r_{\Delta Binary}$ values for individual data sets and the associated causal strength judgments. The dashed black line is the negative bivariate regression. The two solid lines are the regression lines for $r_{\Delta Binary}$ within the two groups of r_{States}, which indicate positive effects of $r_{\Delta Binary}$.

Reverse-coding analysis. An alternative way to control for r_{States} is to recode stimuli in the $r_{States} = -.83$ condition back into the positive domain by flipping the X values around the midpoint. Within each dataset, transitions that were positive (negative) are flipped to become negative (positive), so the model predictions for $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ are also reversed. In Figure 12A, this amounts to transposing stimuli from Regions 7–12 to their reversed counterparts in Regions 1–6.

Because all three predictors were reverse-coded in the $r_{States} = -.83$ condition, we also reversed the causal strength judgments (as if all judgments were made for stimuli with $r_{States} = .83$). Reverse-coding the data meant that all stimuli had $r_{States} = .83$, so there was no need to control for r_{States} statistically. This was beneficial because in the raw-coding analysis, r_{States} and $r_{\Delta Continuous}$ were strongly positively correlated.

We used the same regression models as the raw-coding analysis, except that we dropped the r_{States} predictor and its random slope. The results for this analysis are presented in Table 8. In bivariate regressions, both $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ were significantly positively correlated with the causal strength judgments. In the multivariate analysis including both $r_{\Delta Continuous}$ and $r_{\Delta Binary}$, only $r_{\Delta Binary}$ was significant.

Discussion

The findings suggest that when judging causal strength, participants made greater use of the direction of change in the variables than the magnitude of change (i.e., they were sensitive to $r_{\Delta Binary}$ but not $r_{\Delta Continuous}$), controlling for r_{States}.

There are three drawbacks to the design of Experiment 4. The first is that using the circle for the effect allows the possibility that participants could focus on the diameter or area; however, ancillary analyses show that the results are consistent in either case.[6] The second is that the data sets involve a somewhat unusual case of Simpson's paradox. The third is that the bivariate effect of

[6] We reanalyzed all the data using the area of the circle instead of diameter to calculate the three predictors; the pattern of results remained unchanged. In all but two cases, the *p*-values that were significant remained significant, and *p*-values that were not significant remained nonsignificant. The coefficient estimates and effect sizes were also similar. In the raw-coding analysis, the effect of $r_{\Delta Continuous}$ controlling for r_{States} changed from marginally significant ($p = .045$ in Table 7) to nonsignificant ($B = .85$, $SE = .49$, $p = .08$, partial-$R^2 = .006$, $d = .16$). In the reverse-coding analysis, the bivariate regression of $r_{\Delta Continuous}$ changed from marginally significant ($p = .03$ in Table 8) to nonsignificant ($B = .85$, $SE = .46$, $p = .07$, $r^2 = .006$, $d = .16$). In sum, the main take-home message of Experiment 4, that r_{States} and $r_{\Delta Binary}$ but not $r_{\Delta Continuous}$ explain additional variance in participants' judgments, is consistent regardless of whether participants focus on the area or diameter of the circle. We did not do this reanalysis in Experiments 1A, 1B, 2, and 3. In those experiments, the positive versus negative r_{States} conditions would remain positive and negative, regardless of whether they are calculated with area or diameter. Furthermore, in those experiments, we do not distinguish between $r_{\Delta Continuous}$ and $r_{\Delta Binary}$, and the positive vs. negative transitions conditions would remain positive or negative regardless of whether they are calculated with area or diameter, so there was no need to redo the analysis.

Table 7
Results for Regressions in Experiments 4 and 5 (Raw Data)

Model	Experiment	Predictor											
		r_{States}				$r_{\Delta Continuous}$				$r_{\Delta Binary}$			
		B (*SE*)	*p*	r^2	*d*	*B* (*SE*)	*p*	r^2	*d*	*B* (*SE*)	*p*	r^2	*d*
Bivariate	4	3.89 (.33)	<.001	.40	1.63	5.22 (.43)	<.001	.29	1.28	−2.52 (.56)	<.001	.03	.35
	5	4.04 (.25)	<.001	.27	1.22	3.91 (.25)	<.001	.27	1.22	4.32 (.30)	<.001	.25	1.15
Controlling for r_{States}													
$r_{States} + r_{\Delta Continuous}$	4	3.37 (.44)	<.001	.15	.84	1.00 (.49)	.045	.007	.17	—	—	—	—
	5	1.73 (.50)	<.001	.01	.20	2.35 (.53)	<.001	.01	.20	—	—	—	—
$r_{States} + r_{\Delta Binary}$	4	4.41 (.34)	<.001	.40	1.63	—	—	—	—	2.41 (.48)	<.001	.04	.41
	5	2.52 (.27)	<.001	.06	.51	—	—	—	—	2.13 (.35)	<.001	.03	.35
Multivariate: $r_{States} + r_{\Delta Continuous} + r_{\Delta Binary}$	4	4.90 (.50)	<.001	.15	.84	−.75 (.58)	.19	.002	.09	2.86 (.60)	<.001	.03	.35
	5	1.82 (.50)	<.001	.01	.20	.87 (.54)	.11	.001	.06	1.90 (.37)	<.001	.02	.29

Note. In models with multiple predictors, r^2 represents partial-r^2.

Figure 13. Participants' causal strength judgments by $r_{\Delta Binary}$ values in Experiment 4. There is a negative bivariate relationship (dashed black line). However, controlling for the r_{States} value, there is a positive relationship (solid black or gray lines for each r_{States} group).

$r_{\Delta Binary}$ on the causal strength judgments was negative in the raw-coding analysis. This negative bivariate relation is explained by the negative correlation between the r_{States} and $r_{\Delta Binary}$ values of the stimuli, but it would be ideal to have a positive bivariate effect of $r_{\Delta Binary}$ to increase the confidence of the influence of $r_{\Delta Binary}$ on participants' causal strength judgments. In the final experiment, we sought to replicate the findings here using data sets created with a more typical generative process.

Experiment 5

Method

Participants. One-hundred participants were recruited using Amazon MTurk and paid $1.50. The experiment consisted of 16 scenarios and lasted about 10 min–12 min in total. An additional 10 participants dropped out before completing all scenarios in the experiment (contributing an additional 35 scenarios); we included their partial data in the analysis.

Design and stimuli. We generated 2,000 data sets with the following parameters, using the *mvrnorm* function from the R package *MASS*. Each dataset had 10 observations; we used a smaller number of observations than in previous studies because simulations showed that more observations resulted in a stronger correlation between $r_{\Delta Continuous}$ and $r_{\Delta Binary}$. X and Y were randomly sampled from Gaussian distributions with means of 50 and standard deviations of 25, and a correlation such that r_{States} was exactly .50. Only data sets with X and Y values within the range of 1–100 were used in the study. For each dataset, we created a reversed version with $r_{States} = -.50$ by flipping the values of X around the midpoint. We then calculated the $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ values of the data sets, displayed in Figure 14. Each participant was presented with 16 data sets, randomly sampled from both the $r_{States} = .50$ and $-.50$ versions.

Because X and Y were randomly sampled, the autocorrelations were close to zero. The average autocorrelations within both the $r_{States} = .50$ and $-.50$ versions of data sets were slightly negative

Table 8
Results for Regressions in Experiments 4 and 5 (Reverse-Coded Data)

Model	Experiment	Predictor: $r_{\Delta Continuous}$ B (SE)	p	r^2	d	Predictor: $r_{\Delta Binary}$ B (SE)	p	r^2	d
Bivariate	4	1.00 (.47)	.03	.007	.17	2.41 (.48)	<.001	.04	.41
	5	2.18 (.53)	<.001	.01	.20	2.05 (.34)	<.001	.03	.35
Multivariate: $r_{\Delta Continuous} + r_{\Delta Binary}$	4	−.76 (.57)	.19	.002	.09	2.87 (.59)	<.001	.03	.35
	5	.75 (.54)	.17	.001	.06	1.84 (.36)	<.001	.02	.29

Note. In models with multiple predictors, r^2 represents partial-r^2.

($M = -.12$, $SD = 0.30$ for both X and Y). (This is a general feature of autocorrelation, not something unique to our data; we confirmed through simulation that the sampling distribution of the autocorrelation function of a limited number of randomly generated observations is in fact slightly negative).

The correlations between r_{States}, $r_{\Delta Continuous}$, and $r_{\Delta Binary}$ among all data sets viewed by participants including both the $r_{States} = .50$ and $-.50$ versions are presented in Table 6. The correlations between predictors were higher than in Experiment 4. However, after using the reverse-coding procedure, eliminating the need to control for r_{States}, the correlation between $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ was reduced to .52, low enough for the effects to be disentangled.

Procedure. The procedure was largely similar to the prior experiments, except that the concentration of the drug (X) was displayed with a vertical slider, similar to Experiments 2 and 3. The size of the microorganism (Y) was mapped to the diameter of a circle, similar to Experiments 1 and 4 (see Figure 15). The shade and color of the circle remained constant throughout each scenario.

Results

The results were analyzed in the same way as Experiment 4.

Raw-coding analysis. The results of the raw-coding analyses are presented in Table 7. In the bivariate regressions, all three predictors significantly predicted causal strength judgments. In the second set of regressions that controlled for r_{States}, $r_{\Delta Continuous}$, and $r_{\Delta Binary}$ were still significant predictors. In the multivariate analysis with all three predictors, $r_{\Delta Binary}$ was significant, but $r_{\Delta Continuous}$ was not.

Reverse-coding analysis. We followed the same reverse-coding procedure from Experiment 4. In the reverse-coding analysis, r_{States} was held constant, so only $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ were included as predictors. The results are reported in Table 8. In the bivariate regressions, both $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ were significant predictors. In the multivariate analysis, $r_{\Delta Binary}$ was significant, but $r_{\Delta Continuous}$ was not.

The results of both the raw and reverse-coding analyses were robust regardless of whether the predictors used the diameter or area of the circle; all significant effects were still significant, and all nonsignificant effects were still nonsignificant.

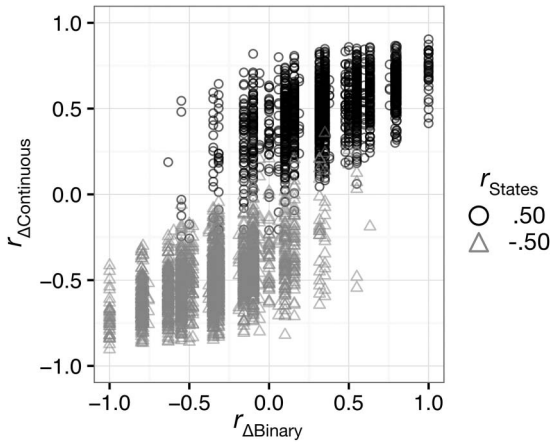

Figure 14. $r_{\Delta Continuous}$ and $r_{\Delta Binary}$ values for 2,000 data sets generated for Experiment 5 and their flipped counterparts. Stimuli viewed by participants were sampled from all data sets shown here.

Discussion

The results of Experiment 5 were consistent with findings from Experiment 4; $r_{\Delta Binary}$, but not $r_{\Delta Continuous}$, was a significant predictor of causal strength judgments over and above the other predictors. In sum, over and above the raw states, participants' causal strength judgments are sensitive mainly to whether variables increase or decrease in a particular transition and not the magnitude of change.

General Discussion

The current research investigated how people learn causal relations in time series settings with multilevel variables. In time series settings, it is critical to control for temporal trends in the data, and one way to do this is to utilize the difference scores in how variables change over time, which we call transitions, rather than the states of the variables at a given instant. Overall, we found that in longitudinal scenarios, in addition to using the overall correlation between the cause and effect, people also used the transitions of how the cause and effect changed from one time point to the next for inferring the strength of the causal relationship.

In Experiments 1A and 1B, we presented participants with data sets in which a cause (X) and an effect (Y) exhibited increasing or decreasing temporal trends. We manipulated the order of observations to create all positive or all negative transitions (varying $r_{\Delta Continuous}$), while holding the correlations between the states of X and Y (r_{States}) constant. The transitions accounted for roughly two to three times more variance in participants' causal strength judg-

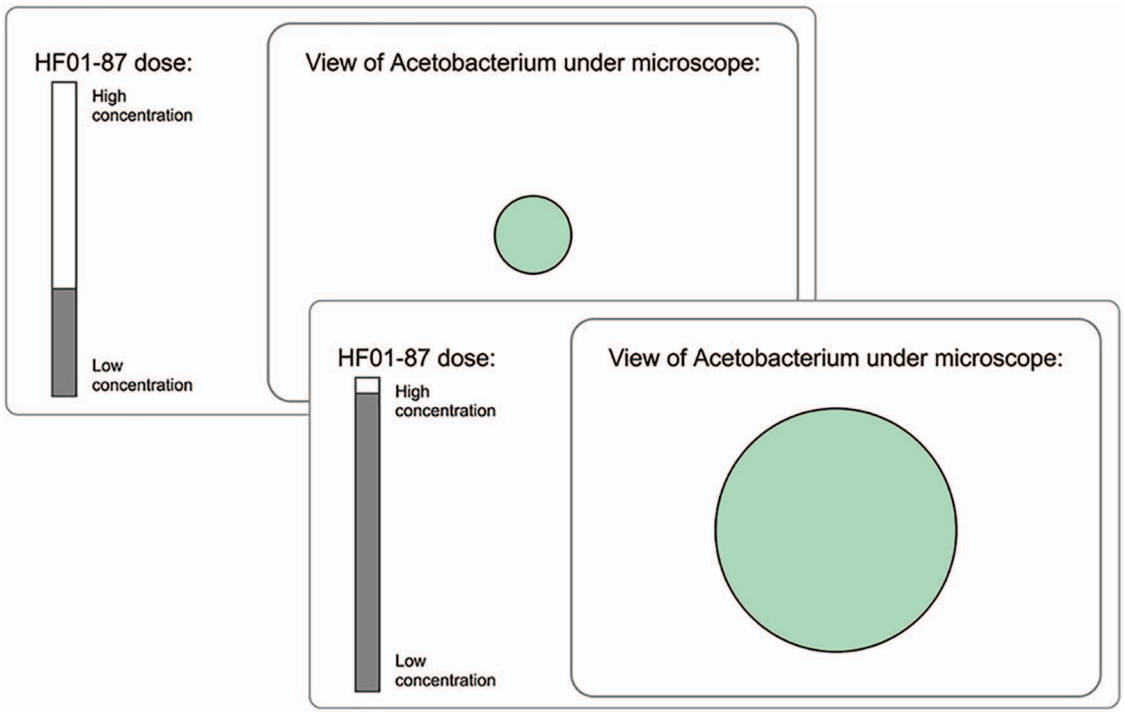

Figure 15. Presentation of stimuli in Experiment 5. Two observations are displayed to show how a transition would appear. See the online article for the color version of this figure.

ments than the states. This effect, if anything, was magnified by the presence of a delay between the cause and the effect.

In Experiment 2, participants relied more on the transitions when stimuli were presented visually than numerically. The numerical condition was the only condition in which the effect size for states was larger than the effect size for transitions. Using transitions helped participants uncover the true causal relation, which raises the following paradox: The ability to precisely represent numbers as symbols enables the performing of complex mathematical operations, but this ability appears to interfere with accurate causal learning in time series contexts.

Experiment 3 ruled out the possibility that instead of relying on transitions, participants were merely sensitive to a primacy or recency effect due to a limited memory of experienced events. There was no main effect of recency, and while we found a small interaction between transitions and recency, there was a large main effect of transitions that cannot be explained by primacy or recency.

In Experiments 4 and 5, we found that in addition to utilizing the states, when participants used transitions they focused simply on whether the cause and effect increased or decreased, not the magnitude of increase or decrease, which could be viewed as a simplifying heuristic for estimating causal strength. It was challenging to discriminate between these two strategies, $r_{\Delta\text{Continuous}}$ and $r_{\Delta\text{Binary}}$, because the most straightforward methods for generating random data results in $r_{\Delta\text{Continuous}}$ and $r_{\Delta\text{Binary}}$ being highly correlated, and because $r_{\Delta\text{Continuous}}$ is also very highly correlated with r_{States}. A number of different techniques were employed to overcome this problem of multicollinearity including (a) holding r_{States} constant in the design of the study rather than controlling for it statistically, (b) using a Simpson's paradox technique to generate the data sets, and (c) reducing the numbers of observations within a dataset. Both experiments found the same results, that after controlling for the other two predictors, r_{States} and $r_{\Delta\text{Binary}}$ were significant, but $r_{\Delta\text{Continuous}}$ was not.

Though the effect sizes of $r_{\Delta\text{Binary}}$ were smaller in the multivariate analyses in Experiments 4 and 5 compared with in Experiments 1 and 2, this was expected given that the goal of Experiments 4 and 5 was to distinguish $r_{\Delta\text{Continuous}}$ and $r_{\Delta\text{Binary}}$, which are typically highly correlated. This meant that the most extreme cases used in Experiments 1 and 2, for which $r_{\Delta\text{Continuous}}$ and $r_{\Delta\text{Binary}}$ converged, could not be used for Experiments 4 and 5. In contrast, in Experiments 4 and 5, there were no linear trends in the data. (When there are linear trends, using difference scores is normatively justified.) Still, the partial variance explained by $r_{\Delta\text{Binary}}$ was in the range of .02 to .04. Converted into Cohen's *d*, this is equivalent to a range of .29 to .41, "small" to "medium" effect sizes by convention. In the multivariate analyses in Experiments 1 and 2, the effect sizes for $r_{\Delta\text{Continuous}}$ (which is essentially equivalent to $r_{\Delta\text{Binary}}$) convert to *d* of .59 to 1.4, which are "medium" to "large" effects.

Using Difference Scores Versus Regression to Control for Temporal Trends

We have proposed a heuristic for causal learning with time series; people attend to transitions to account for temporal trends, akin to taking difference scores to account for nonstationarity in time series data. An alternative way to account for temporal trends in time series data is with a regression model predicting the effect

from the observed cause that includes time as a covariate. This approach could be considered an *ideal observer* model in which the learner "knows" about the presence of temporal trends.

In the simple case of increasing or decreasing linear trends in the data (as in Experiments 1 and 2), an ideal observer model that knows to account for linear trends will reach very similar conclusions to the simplifying heuristics of learning from transitions. Indeed, both of these approaches are viable options for dealing with nonstationary time series (Shumway & Stoffer, 2011). In Appendix A (Part 1), we show how models using first-order difference scores control for linear trends. The fact that participants are more likely to infer a positive causal relation in the positive transitions condition than the negative transitions condition provides evidence that people are able to control for temporal trends to some extent.

However, there are two reasons to believe that participants were using difference scores and were not using an analysis akin to regression controlling for time. First, in Experiment 3, the cause and effect undergo more complex trends; one of the variables undergoes a positive trend in the first half of the data followed by a negative trend. In order to accurately control for the temporal trends without using difference scores, participants would have to control for a nonlinear temporal trend that they do not know about in advance. However, an ideal observer model could theoretically involve more complex functions than the linear model discussed above. Such models could account for cases in which the relationships between time and the variables are nonlinear. Future experiments using data with nonlinear temporal trends will be required to discriminate between this class of ideal observer models and our proposed heuristic.

Second, in Experiments 4 and 5, the trials were randomly ordered such that there were no temporal trends on average, yet we still see an effect of transitions in these experiments. In this case, a standard ideal observer would not include time as a covariate. Perhaps a Gaussian process model that is sensitive to short trends in time (despite no overall trends on average) could also explain this effect.

In sum, while additional work is needed to rule out the possibility that participants are using a more sophisticated strategy, some of our findings suggest that participants are using a simple heuristic of focusing on transitions to account for temporal trends.

Do Learners Use States or Transitions as a Default Tendency?

An important question to be answered is whether people use transitions for inferring causal strength both in situations in which transitions are statistically useful (e.g., nonstationary time series environments), and also in situations in which using transitions are not necessary from a statistical perspective (e.g., cross-sectional environments). Phrased another way, do learners switch between using transitions versus states for different kinds of environments?

From a normative perspective, Appendix A (Part 1) demonstrates how taking the correlation between the difference scores of X and Y effectively removes the confounding effect of temporal trends in nonstationary environments, allowing the true causal relation to be uncovered. A follow-up question is whether using difference scores instead of raw scores results in worse performance in a stationary (independent and identically distributed; *iid*) environment. In Appendix A (Part 2), we demonstrate that in a stationary environment, taking a correlation of difference scores ($r_{\Delta\text{Continuous}}$) does not dramatically affect the precision or bias of the correlation, and $r_{\Delta\text{Binary}}$ also tracks r_{States} monotonically. Collectively, this means that using transitions works pretty well in both stationary and nonstationary environments, whereas using states works well in stationary environments but can produce high degrees of error in nonstationary environments. We do not presume that a learner must choose between these strategies, and indeed we have evidence that they use both in both environments. However, if a learner were forced to choose one strategy for all environments, this analysis suggests that using transitions would be a better choice.

Experiment 5 is especially relevant to answering this question, because in Experiment 5 the data sets were generated randomly with an iid process, and the data sets were sampled randomly from the entire set that was generated. In Experiment 5, the effect of $r_{\Delta\text{Binary}}$ was significant, which suggests that even in iid settings for which transitions are not statistically necessary to reach accurate judgments, that people still do use transitions in addition to states. A similar conclusion was reached in another study in which participants learned causal strength from binary stimuli (Soo & Rottman, 2015). Collectively, these findings raise the possibility that participants may have been using transitions as well as states in previous studies on causal learning that involved a trial-by-trial presentation of randomly ordered data.

Toward a Process-Level Account of Learning Causal Strength From Transitions

So far, our account of the use of transitions has been situated mainly at the computational level of Marr's (1982) hierarchy. The goal for $r_{\Delta\text{Continuous}}$ and $r_{\Delta\text{Binary}}$ relative to r_{States} is to solve the problem that temporal confounds can distort the apparent relation between a cause and an effect. In Experiments 4 and 5, we introduced the $r_{\Delta\text{Binary}}$ model. Though this model is also most appropriately conceived at the computational level, simplifying the data into a binary representation raises the possibility that learners could apply simpler rule-based models like ΔP (Jenkins & Ward, 1965), Power PC (Cheng, 1997), or others (e.g., Hattori & Oaksford, 2007) for calculating causal strength from a 2×2 contingency table.

It is also possible to elaborate an algorithmic or process-level version of $r_{\Delta\text{Continuous}}$ and $r_{\Delta\text{Binary}}$. In many situations, the Rescorla-Wagner learning algorithm (RW; Miller, Barnet, & Grahame, 1995; Rescorla & Wagner, 1972; Wagner & Rescorla, 1972) and the generalized Delta Rule (Rumelhart, Hinton, & Williams, 1986) compute conditional contrasts similar to regression. We built a modified version of RW that instead of taking the raw states, takes the difference scores of the cause and the effect as input, and ran simulations for stimuli from Experiments 1, 2, and 3 (see Appendix B for our method). RW did not learn the correct polarity of causal strength based on the raw states. However, our modified RW did learn the correct polarity based on the difference scores, with predictions approximating participants' judgments. In sum, RW provides one way to implement the strategies described here using a computationally tractable learning process.

Philosophical Perspectives on States Versus Transitions

There are important philosophical issues relevant to causal learning from time series data concerning the metaphysics of the cause, the effect, and relation between the two. Although this topic is very broad, we highlight three specific questions that are relevant to the present research. First, what are the cause and effect (otherwise known as the *relata*)? Many philosophers view causes and effects as *events* (Schaffer, 2016). However, this is not a universal position. For example, Lewis (1973) focuses primarily on events, though allows for the possibility of other sorts of causes and effects, which sometimes include properties or facts (see also Mellor, 1995, 2004). Second, if causation is among events, what constitutes an event? With regards to the current research, the most relevant distinction is whether an event must involve a change (called a *dynamic event*) or not (a *static event*; see Casati & Varzi, 2015). Though some philosophers argue that events must involve changes, we do not believe that this is a common position. For example, Lewis (1973) gives an example of a barometer reading depending on the pressure. In this case, there is a clear cause–effect relationship (between air pressure and the barometer), but the events are simply the states of the air pressure and the barometer reading at a given time. A third question is how to understand *negative events* (the absence of an event), and whether negative events can participate in a causal relation (Casati & Varzi, 2015; Wolff, Barbey, & Hausknecht, 2010)? This is especially problematic if events are viewed to require a change, because then any nonchange cannot be a cause. For example, imagine a table supporting a plate. A strict position that events must comprise changes, and that causal relata must be events, would require arguing that there is no causal relation between the table and the plate. To avoid this sort of problem, we believe most philosophers (and statisticians) hold fairly broad views of "events" and what sorts of events can participate in causal relations.[7] However, there also seems to be a view that changes are an especially important class of events.

Certain philosophers also take a stance that a causal claim is essentially a claim about change over time when tracking an entity longitudinally. For example, Woodward (2003) argues that when we say that "X causes Y," what we really mean is that if we were able to change the value of X over time, the value of Y would change. Woodward (2003) argues that this is what we mean when making a causal claim even if the causal claim is based on cross-sectional data (e.g., from a randomized-controlled study) that does not allow for a change (difference) score analysis.

Open Questions

There are a number of open questions and future directions. First, in the present article, we focused on the case of elemental causal induction for which the goal was to learn about causal strength when there was a single cause and a single effect. However, in the real world typically multiple causes combine or interact to produce an effect (Novick & Cheng, 2004; Spellman, 1996; Waldmann & Holyoak, 1992). In current work, we are investigating how people learn about multiple causes of a single effect in a time series setting (Derringer & Rottman, 2016).

Second, what remaining factors influence whether a learner controls for temporal trends? In Experiment 2, we found that participants were more likely to control for temporal trends when the stimuli were presented visually rather than numerically. However, even in the visual condition, participants still did not give extremely high (or low) judgments in the positive (negative) transitions conditions, and participants still used states to some extent. It is possible that participants' explicit beliefs about background causes such as time may moderate the extent to which learners control for temporal trends by using transitions. For example, Gureckis and Love (2009) demonstrated that when people are shown a cue representing an underlying state in the environment, they are better at learning the payoffs of two options. If people are cued in to a variable that predicts temporal trends, learners may use transitions even more as they seek to control for the nonstationarity.

A third question is whether the process of causal induction studied here for multilevel variables can be generalized to cases where people learn about a binary cause and a binary effect. The majority of research on causal learning has focused on binary variables, and our focus on continuous variables was a deliberate choice—we wanted to study temporal trends that could increase or decrease across many values (e.g., Figure 1). Binary variables can exhibit autocorrelation—a variable can be "off" for multiple observations before being "on" for multiple observations—but they cannot exhibit the same dramatic nonstationary increasing or decreasing trends like in Figure 1A. We are currently studying whether there are similar effects of transitions even for binary variables (Soo & Rottman, 2015).

A fourth question is how memory constraints influence people's reliance on transitions and reliance on $r_{\Delta\text{Binary}}$. One benefit of discretizing difference scores ($r_{\Delta\text{Binary}}$) from an algorithmic perspective is that it reduces the memory demands of learning relative to states or continuous difference scores. Although our experiments used trial-by-trial paradigms with only 10–20 observations, it is possible that if the memory load were further reduced by using even shorter time series, or by decreasing the range of possible levels for the cause and effect, that there would be a shift toward using states.

Lastly, in the introduction we cited a variety of research on the importance of covariation detection for many areas of cognition including categorization, stereotype formation, as well as its role in a variety of clinical disorders including depression, phobias, and schizophrenia. That research differs from the current studies in that it has primarily focused on covariation detection rather than causal strength induction—the difference being that for causal strength induction one variable is a putative cause and the other a putative effect. Further, most of that work has focused on atemporal cases (but see Sakamoto, Jones, & Love, 2008, for an example of category learning involving temporal trends). Given the similarity between the tasks involved in causal strength induction, covariation detection, and category learning, an important future direction is investigating whether people use processes similar to the one proposed here for causal strength induction when performing other tasks.

[7] A personal communication with James Woodward supports our assessment. We also thank him for the table-plate example.

Conclusion

Six experiments provided evidence that people use the transitions of how a cause and effect change over time to infer whether the cause has a positive or negative influence on the effect from time series data. Most research on causal strength learning has focused on situations in which the order of the trials is random, whereas here we focused on time series contexts, which add an additional layer of complexity for inferring causal strength. These experiments provide an optimistic view of human causal learning in longitudinal contexts in that participants were able to uncover the strengths of causal relationships despite temporal trends that can obscure the true causal relation.

Context of the Research

The current studies are part of a larger research program aiming to understand causal learning in nonstationary environments. The core idea for this project arose a number of years ago while the second author was studying how people learn about tolerance and sensitization effects (Rottman & Ahn, 2009). For example, caffeine initially has a very strong effect on attention, but with repeated usage has diminishing effects. The similarity between that research and the current research is that both involve causal learning about variables that are continuous or at least multilevel rather than binary, and in both, the simple correlation (r_{States}) provides little insight into understanding the causal mechanism whereas the changes do provide insight. The broader hypothesis is that people are often able to learn how to control moderately dynamically complex causal systems, and the empirical question is how people are able to do so, and in particular, if there are relatively simple heuristics that can adaptively guide behavior. We are currently extending this research to understand if and how similar principles of attending to changes also play a role when learning about binary variables (Soo & Rottman, 2015), as well as to understand how well different reinforcement-learning models perform in these nonstationary settings. More broadly, we hope to understand how people utilize different learning processes in different environments, and how adaptively people can switch between them in different environments.

References

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108,* 441–485. http://dx.doi.org/10.1037/0096-3445.108.4.441

Balzan, R. P., Delfabbro, P. H., Galletly, C. A., & Woodward, T. S. (2013). Illusory correlations and control across the psychosis continuum: The contribution of hypersalient evidence-hypothesis matches. *Journal of Nervous and Mental Disease, 201,* 319–327. http://dx.doi.org/10.1097/NMD.0b013e318288e229

Beach, L. R., & Scopp, T. S. (1966). Inferences about correlations. *Psychonomic Science, 6,* 253–254. http://dx.doi.org/10.3758/BF03328053

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica, 87,* 137–154. http://dx.doi.org/10.1016/0001-6918(94)90048-5

Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115,* 396–425. http://dx.doi.org/10.1037/0033-295X.115.2.396

Buehner, M. J. (2005). Contiguity and covariation in human causal inference. *Learning & Behavior, 33,* 230–238. http://dx.doi.org/10.3758/BF03196065

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1119–1140. http://dx.doi.org/10.1037/0278-7393.29.6.1119

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology, 56,* 865–890. http://dx.doi.org/10.1080/02724980244000675

Buehner, M. J., & May, J. (2009). Causal induction from continuous event streams: Evidence for delay-induced attribution shifts. *The Journal of Problem Solving, 2,* 42–80. http://dx.doi.org/10.7771/1932-6246.1057

Burns, P., & McCormack, T. (2009). Temporal information and children's and adults' causal inferences. *Thinking & Reasoning, 15,* 167–196. http://dx.doi.org/10.1080/13546780902743609

Casati, R., & Varzi, A. (2015). Events. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2016 Edition). Retrieved from https://plato.stanford.edu/archives/win2015/entries/events/

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

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365–382. http://dx.doi.org/10.1037/0033-295X.99.2.365

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

Crocker, J. (1981). Judgement of covariation by social perceivers. *Psychological Bulletin, 90,* 272–292. http://dx.doi.org/10.1037/0033-2909.90.2.272

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

Derringer, C., & Rottman, B. M. (2016). Temporal causal strength learning with multiple causes. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 758–763). Austin, TX: Cognitive Science Society.

Díez-Alegría, C., Vázquez, C., & Hernández-Lloreda, M. J. (2008). Covariation assessment for neutral and emotional verbal stimuli in paranoid delusions. *British Journal of Clinical Psychology, 47,* 427–437. http://dx.doi.org/10.1348/014466508X332819

Donkin, C., Rae, B., Heathcote, A., & Brown, S. D. (2015). Why is accurately labelling simple magnitudes so hard? A past, present and future look at simple perceptual judgment. In J. R. Busemeyer, Z. Wang, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 121–141). New York, NY: Oxford University Press.

Erlick, D. E. (1966). Human estimates of statistical relatedness. *Psychonomic Science, 5,* 365–366. http://dx.doi.org/10.3758/BF03328441

Erlick, D. E., & Mills, R. G. (1967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology, 73,* 9–14. http://dx.doi.org/10.1037/h0024138

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8,* 307–314. http://dx.doi.org/10.1016/j.tics.2004.05.002

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 678–693. http://dx.doi.org/10.1037/a0014928

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General, 141,* 124–133. http://dx.doi.org/10.1037/a0024006

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. http://dx.doi.org/10.3758/MC.36.6.1087

Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General, 139,* 756–771. http://dx.doi.org/10.1037/a0020976

Greville, W. J., Cassar, A. A., Johansen, M. K., & Buehner, M. J. (2013). Structural awareness mitigates the effect of delay in human causal learning. *Memory & Cognition, 41,* 904–916. http://dx.doi.org/10.3758/s13421-013-0308-7

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

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition, 113,* 293–313. http://dx.doi.org/10.1016/j.cognition.2009.03.013

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30,* 1128–1137. http://dx.doi.org/10.3758/BF03194330

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science, 31,* 765–814. http://dx.doi.org/10.1080/03640210701530755

Helson, H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review, 55,* 297–313. http://dx.doi.org/10.1037/h0056721

Helson, H. (1964). Current trends and issues in adaptation-level theory. *American Psychologist, 19,* 26–38. http://dx.doi.org/10.1037/h0040013

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62,* 135–163. http://dx.doi.org/10.1146/annurev.psych.121208.131634

Huq, S. F., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology, 40,* 801–812. http://dx.doi.org/10.1080/14640748808402300

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 79,* 1–17. http://dx.doi.org/10.1037/h0093874

Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiology, 16,* 85–125.

Krantz, D. L., & Campbell, D. T. (1961). Separating perceptual and linguistic effects of context shifts upon absolute judgments. *Journal of Experimental Psychology, 62,* 35–42. http://dx.doi.org/10.1037/h0040386

Kutzner, F. L., & Fiedler, K. (2015). No correlation, no evidence for attention shift in category learning: Different mechanisms behind illusory correlations and the inverse base-rate effect. *Journal of Experimental Psychology: General, 144,* 58–75. http://dx.doi.org/10.1037/a0038462

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 856–876. http://dx.doi.org/10.1037/0278-7393.30.4.856

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 451–460. http://dx.doi.org/10.1037/0278-7393.32.3.451

Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal, 3,* 184–195.

Lane, D. M., Anderson, C. A., & Kellam, K. L. (1985). Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 640–649. http://dx.doi.org/10.1037/0096-1523.11.5.640

Le Pelley, M. E., Reimers, S. J., Calvini, G., Spears, R., Beesley, T., & Murphy, R. A. (2010). Stereotype formation: Biased by association. *Journal of Experimental Psychology: General, 139,* 138–161. http://dx.doi.org/10.1037/a0018210

Lewis, D. (1973). Causation. *The Journal of Philosophy, 70,* 556–567. http://dx.doi.org/10.2307/2025310

Marr, D. (1982). *Vision: A computational investigation into human representation and processing of visual information*. San Diego, CA: Freeman.

Marsh, J. K., & Ahn, W.-K. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 334–352. http://dx.doi.org/10.1037/a0014929

McCormack, T., Frosch, C., Patrick, F., & Lagnado, D. (2015). Temporal and statistical information in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 395–416. http://dx.doi.org/10.1037/a0038385

Mellor, D. H. (1995). *The facts of causation*. London, UK: Routledge. http://dx.doi.org/10.4324/9780203302682

Mellor, D. H. (2004). For facts as causes and effects. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 309–323). Cambridge, MA: MIT Press.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin, 117,* 363–386. http://dx.doi.org/10.1037/0033-2909.117.3.363

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111,* 455–485. http://dx.doi.org/10.1037/0033-295X.111.2.455

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive t tests: Lay use of statistical information. *Psychonomic Bulletin & Review, 14,* 1147–1152. http://dx.doi.org/10.3758/BF03193104

Ohman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108,* 483–522. http://dx.doi.org/10.1037/0033-295X.108.3.483

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review, 14,* 577–596. http://dx.doi.org/10.3758/BF03196807

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. E. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Restle, F., & Merr, C. T. (1968). An adaptation-level theory account of a relative-size illusion. *Psychonomic Science, 12,* 229–230. http://dx.doi.org/10.3758/BF03331284

Rottman, B. M. (2016). Searching for the best cause: Roles of mechanism beliefs, autocorrelation, and exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 1233–1256. http://dx.doi.org/10.1037/xlm0000244

Rottman, B. M., & Ahn, W.-K. (2009). Causal learning about tolerance and sensitization. *Psychonomic Bulletin & Review, 16,* 1043–1049. http://dx.doi.org/10.3758/PBR.16.6.1043

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology, 64*(1–2), 93–125. http://dx.doi.org/10.1016/j.cogpsych.2011.10.003

Rottman, B. M., Kominsky, J. F., & Keil, F. C. (2014). Children use temporal cues to learn causal directionality. *Cognitive Science, 38,* 489–513. http://dx.doi.org/10.1111/cogs.12070

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the*

microstructure of cognition, vol. 1: Foundations (pp. 318–362). Cambridge, MA: MIT Press.

Saito, M. (2015). How people estimate effect sizes: The role of means and standard deviations. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2075–2079). Austin, TX: Cognitive Science Society.

Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory & Cognition, 36,* 1057–1065. http://dx.doi.org/10.3758/MC.36.6.1057

Sarris, V. (1967). Adaptation-level theory: Two critical experiments on Helson's weighted-average model. *The American Journal of Psychology, 80,* 331–344. http://dx.doi.org/10.2307/1420364

Schaffer, J. (2016). The metaphysics of causation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2016 Edition). Retrieved from https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology, 41B,* 139–159.

Sherman, J. W., Kruschke, J. K., Sherman, S. J., Percy, E. J., Petrocelli, J. V., & Conrey, F. R. (2009). Attentional processes in stereotype formation: A common model for category accentuation and illusory correlation. *Journal of Personality and Social Psychology, 96,* 305–323. http://dx.doi.org/10.1037/a0013778

Shumway, R. H., & Stoffer, D. S. (2011). *Time series analysis and its applications*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-7865-3

Soo, K. W., & Rottman, B. M. (2014). Learning causal direction from transitions with continuous and noisy variables. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1485–1490). Austin, TX: Cognitive Science Society.

Soo, K. W., & Rottman, B. M. (2015). Elemental causal learning from transitions. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2254–2259). Austin, TX: Cognitive Science Society.

Soo, K. W., & Rottman, B. M. (2016). Causal learning with continuous variables over time. In A. Papafrogou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 153–158). Austin, TX: Cognitive Science Society.

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

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112,* 881–911. http://dx.doi.org/10.1037/0033-295X.112.4.881

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53,* 1–26. http://dx.doi.org/10.1016/j.cogpsych.2005.10.003

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations* (pp. 444–459). Cambridge, MA: MIT Press.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review, 34,* 273–286. http://dx.doi.org/10.1037/h0070288

Thurstone, L. L. (1927b). Psychophysical analysis. *The American Journal of Psychology, 38,* 368–389. http://dx.doi.org/10.2307/1415006

Vigen, T. (2015). *Spurious correlations*. New York, NY: Hachette Book Group.

Vogel, T., Kutzner, F., Freytag, P., & Fiedler, K. (2014). Inferring correlations: From exemplars to categories. *Psychonomic Bulletin & Review, 21,* 1316–1322. http://dx.doi.org/10.3758/s13423-014-0586-5

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and learning* (pp. 301–336). London, UK: Academic Press.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121,* 222–236. http://dx.doi.org/10.1037/0096-3445.121.2.222

White, P. A. (2015). Causal judgements about temporal sequences of events in single individuals. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 68,* 2149–2174. http://dx.doi.org/10.1080/17470218.2015.1009475

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General, 139,* 191–221. http://dx.doi.org/10.1037/a0018129

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York, NY: Oxford University Press.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society, 89,* 1–63. http://dx.doi.org/10.2307/2341482

(*Appendices follow*)

Appendix A

Estimating Causal Strength From Difference Scores

Part 1: How $r_{\Delta Continuous}$ Uncovers the True Causal Strength With Large Samples in Nonstationary Environments With a Temporal Confound

This section provides a more thorough justification for how the difference score analysis proposed in the main text accurately uncovers the true correlation between a cause X and an effect Y that both increase over time through an unobserved temporal confound. In Figure A1, the generative process from Figure 2 in the main text is elaborated to show the error terms for the two observed variables.

The following two equations represent the data generating process. t is a vector representing time in integers. x_t and y_t are the variable vectors observed by the learner. ε_{xt} and ε_{yt} are vectors of random variables that represent the unique variance in x_t and y_t, respectively.

$$x_t = t + \varepsilon_{xt}, \text{ where } \varepsilon_{xt} \sim N(0, s^2)$$

$$y_t = t + rx_t + (1 - r^2)^{1/2}\varepsilon_{yt}, \text{ where } \varepsilon_{yt} \sim N(0, s^2)$$

By substitution,

$$y_t = (1 + r)t + r\varepsilon_{xt} + (1 - r^2)^{1/2}\varepsilon_{yt}$$

r is a constant between -1 and $+1$ representing the causal strength as a correlation coefficient of ε_{xt} on y_t after accounting for t. The weights involving r on ε_{xt} and ε_{yt} for y_t guarantee that r is the correlation between ε_{xt} and y_t after accounting for t in large samples. This can be demonstrated by showing that the term ε_{xt} accounts for r^2 percent of the variance of y after removing the variance from t:

$$(r\varepsilon_{xt} + (1 - r^2)^{1/2}\varepsilon_{yt}).$$

This is proved in the following two lines:

$$\mathrm{Var}(r\varepsilon_{xt}) = r^2\mathrm{Var}(\varepsilon_{xt}) = r^2s^2$$

$$\begin{aligned}\mathrm{Var}(r\varepsilon_{xt} + (1 - r^2)^{1/2}\varepsilon_{yt}) &= r^2\mathrm{Var}(\varepsilon_{xt}) + (1 - r^2)\mathrm{Var}(\varepsilon_{yt}) \\ &= r^2s^2 + (1 - r^2)s^2 \\ &= s^2\end{aligned}$$

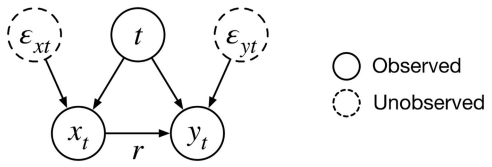

Figure A1. Generative Process From Figure 2 Including Error Terms.

The main problem is that, because ε_{xt} is unobserved, r cannot be directly calculated. The main point of this section is to show that the causal strength r can be uncovered in large samples by taking a correlation of the difference scores of x_t and y_t, which are observable. We show this by demonstrating that the variance contained in the difference scores of x_t, is r^2 percent of the variance in the difference scores of y_t.

$$\begin{aligned}&\mathrm{Var}(\mathrm{Diff}(x_t)) \\ &= \mathrm{Var}(\mathrm{Diff}(t + \varepsilon_{xt})) \\ &= \mathrm{Var}(\mathrm{Diff}(t)) + \mathrm{Var}(\mathrm{Diff}(\varepsilon_{xt}))\end{aligned}$$

Because t is a vector that increases linearly with time [0, 1, 2, 3 . . .], the difference scores of t are [1, 1, 1 . . .], which means that $\mathrm{Var}(\mathrm{Diff}(t)) = 0$, so

$$\mathrm{Var}(\mathrm{Diff}(x_t)) = \mathrm{Var}(\mathrm{Diff}(\varepsilon_{xt}))$$

Through the properties already explained above, it can be shown that

$$\begin{aligned}&\mathrm{Var}(\mathrm{Diff}(y_t)) \\ &= \mathrm{Var}(\mathrm{Diff}((1 + r)t + r\varepsilon_{xt} + (1 - r^2)^{1/2}\varepsilon_{yt}) \\ &= \mathrm{Var}(\mathrm{Diff}((1 + r)t)) + \mathrm{Var}(\mathrm{Diff}(r\varepsilon_{xt})) + \mathrm{Var}(\mathrm{Diff}(1 - r^2)^{1/2}\varepsilon_{yt}) \\ &= 0 + r^2\mathrm{Var}(\mathrm{Diff}(\varepsilon_{xt})) + (1 - r^2)\mathrm{Var}(\mathrm{Diff}(\varepsilon_{yt}))\end{aligned}$$

Finally, because $\mathrm{Var}(\varepsilon_{xt}) = \mathrm{Var}(\varepsilon_{yt})$, this last line shows that $\mathrm{Diff}(\varepsilon_{xt})$ contributes r^2 percent of the variance to $\mathrm{Diff}(y_t)$. And because $\mathrm{Var}(\mathrm{Diff}(x_t)) = \mathrm{Var}(\mathrm{Diff}(\varepsilon_{xt}))$, $\mathrm{Diff}(x_t)$ contributes r^2 percent of the variance to $\mathrm{Diff}(y_t)$.

Part 2: How $r_{States} = r_{\Delta}$ With Large Samples in Stationary Environments (No Temporal Confound)

This section proves that in large samples, when x and y are independent and identically distributed (iid), $\mathrm{cor}(x, y) = \mathrm{cor}(\mathrm{diff}(x), \mathrm{diff}(y))$. Using the terminology from the paper, this means that $r_{States} = r_{\Delta Continuous}$. Using the same logic from Part 1, we set up y_t such that r is the correlation between x_t and y_t, so x_t accounts for r^2 percent of the variance in y_t.

$$x_t \sim N(0, s^2)$$

$$y_t = rx_t + (1 - r^2)^{1/2}\varepsilon_t, \text{ where } \varepsilon_t \sim N(0, s^2)$$

$$\begin{aligned}&\mathrm{Var}(\mathrm{Diff}(y_t)) \\ &= \mathrm{Var}(\mathrm{Diff}(rx_t + (1 - r^2)^{1/2}\varepsilon_t) \\ &= \mathrm{Var}(\mathrm{Diff}(rx_t)) + \mathrm{Var}(\mathrm{Diff}((1 - r^2)^{1/2}\varepsilon_t)) \\ &= r^2\mathrm{Var}(\mathrm{Diff}(x_t)) + (1 - r^2)\mathrm{Var}(\mathrm{Diff}(\varepsilon_t))\end{aligned}$$

(*Appendices continue*)

Finally, because $\mathrm{Var}(x_t) = \mathrm{Var}(\varepsilon_t)$, the last line in the equation above shows that $\mathrm{Diff}(x_t)$ accounts for r^2 percent of the variance in $\mathrm{Diff}(y_t)$.

We also ran two simulations to estimate the means (bias) and standard deviations (precision) of the estimate of cor(x, y) versus cor(diff(x), diff(y)), or r_{States} versus $r_{\Delta\mathrm{Continuous}}$. We also compared these with $r_{\Delta\mathrm{Binary}}$. The first simulation, summarized in Table A1, computed the mean value of r_{States}, $r_{\Delta\mathrm{Continuous}}$, and $r_{\Delta\mathrm{Binary}}$ for 10,000 simulated data sets with a sample size of 1,000 observations each. Seven versions were run with different degrees of correlation between X and Y, from 0 to 1, with a step size of one sixth. The data generation process used the same process explained in Part 1. The overall conclusion is that with a large number of observations, r_{States} and $r_{\Delta\mathrm{Continuous}}$ are virtually identical. The mean value for $r_{\Delta\mathrm{Binary}}$ is slightly lower, but follows the same increasing trend.

The second simulation measured the precision of r_{States}, $r_{\Delta\mathrm{Continuous}}$, and $r_{\Delta\mathrm{Binary}}$ as estimates of the correlation r. For sample sizes between 10 and 1,000, we created 10,000 data sets with randomly generated noise using the same data generating process as in Part 1 with $r = .7071$ (i.e., $r^2 = .50$). According to this process, r does not exactly equal .7071, but r approaches .7071 for large data sets.

The standard deviations of the estimators are reported in Table A2. The standard deviations follow the pattern $r_{\mathrm{States}} < r_{\Delta\mathrm{Continuous}} < r_{\Delta\mathrm{Binary}}$. $r_{\Delta\mathrm{Continuous}}$ is only slightly worse than r_{States} implying that taking a difference score per se does not dramatically decrease the precision. $r_{\Delta\mathrm{Binary}}$ does have considerably worse precision with small sample sizes. This is not surprising given that it discards useful data. In summary, both $r_{\Delta\mathrm{Continuous}}$ and $r_{\Delta\mathrm{Binary}}$ approximate r_{States}, though the approximation is not perfect.

Table A1
Simulation Results Showing Means of Estimators of r *in a Stationarity Environment (Sample Size = 1,000) With Different Levels of* r

r	0	1/6	1/3	1/2	2/3	5/6	1
r_{States}	0	.166	.333	.500	.666	.833	1
$r_{\Delta\mathrm{Continuous}}$	0	.166	.333	.500	.666	.833	1
$r_{\Delta\mathrm{Binary}}$	0	.106	.216	.333	.464	.627	1

Table A2
Simulation Results Showing Standard Deviations of Estimators of r *in a Stationary Environment* ($r^2 = .5$) *With Different Sample Sizes*

Sample size	10	20	50	100	200	500	1,000
r_{States}	.20	.12	.07	.05	.04	.02	.02
$r_{\Delta\mathrm{Continuous}}$	.24	.15	.09	.06	.04	.03	.02
$r_{\Delta\mathrm{Binary}}$	.31	.21	.13	.09	.06	.04	.03

Appendix B

Implementation of the Rescorla-Wagner Learning Algorithm

In this section, we describe the implementation of the Rescorla-Wagner learning algorithm (RW; Miller et al., 1995; Rescorla & Wagner, 1972; Wagner & Rescorla, 1972) to model predictions for stimuli in Experiments 1, 2, and 3. RW takes as input values of the cause and effect and updates the strength of the cause based on how well its presence predicts the effect. The change in strength for the cause (Δw) after each trial is computed as follows: $\Delta w_{t+1} = \alpha \cdot (y_t - w_t x_t) \cdot x_t$

α is a learning rate parameter, which we set to .10. x_t and y_t represent the current state of the cause and effect. $w_t x_t$ is the prediction of y_t. It is standard practice to include an ever-present background cue, so $w_t x_t$ is the dot product (sum of the products) of the weights of the two cues (the background cue and the cause) and the states of the two cues. The term in the parentheses is the error; the difference between y_t and the prediction of y_t. The rightmost term of x_t scales the amount the weight gets updated by the magnitude of the cause, which is standard practice when the cues are continuous rather than binary (Stone, 1986).

We ran simulations for stimuli using transformed scores of X and Y divided by 100. We did this because when we used the 0–100 scale, the value for w "exploded" and alternated between extremely large positive and negative values.

(*Appendices continue*)

RW Simulations of Stimuli From Experiments 1–3

Figure B1 plots RW's simulated final *w* values for each of the 20 data sets from Experiment 1 and 2. In Experiment 1, the order of the trials was presented in only the forward order, but in Experiment 2 it was presented in both the forward and reverse orders. This simulation shows that RW clearly discriminates between the stimuli in which r_{States} is positive versus negative, but does not reliably discriminate between the positive, negative, and random transitions conditions. The positive transitions condition has higher *w* values than the negative and random transitions conditions in the forward order; this could be due to the fact that RW is sensitive to a recency effect. However, RW predicts lower causal strengths in the positive transition condition when the stimuli order is reversed. In contrast, our participants did not show a difference between the forward versus reversed orderings. RW also fails to capture differences between the negative versus random transition conditions. In sum, RW fails to capture the main trends in the data, especially in the reverse condition.

Figure B2 shows the simulations of RW for the 16 data sets in Experiment 3. RW fails to discriminate between the positive versus negative transition conditions.

Modified Version of RW With Transitions (RWΔ)

We created a modified version of RW, named RWΔ, which runs RW based on the difference scores (also divided by 100). Figure B3 shows the simulation of RWΔ for Experiments 1 and 2; RWΔ captures the trends in the data in Figures 5 and 8 much better than RW. The predictions of RWΔ are very similar to the predictions of $r_{\Delta Continuous}$ and $r_{\Delta Binary}$.

Figure B4 shows the predictions of RWΔ for Experiment 3. The predictions very closely line up with participants' inferences (Figure 10A) as well as with $r_{\Delta Continuous}$ and $r_{\Delta Binary}$.

In sum, while RW is unable to account for our participants' judgments, RWΔ captures their judgments quite well. These simulations show that RWΔ is a computationally tractable way of learning from transitions.

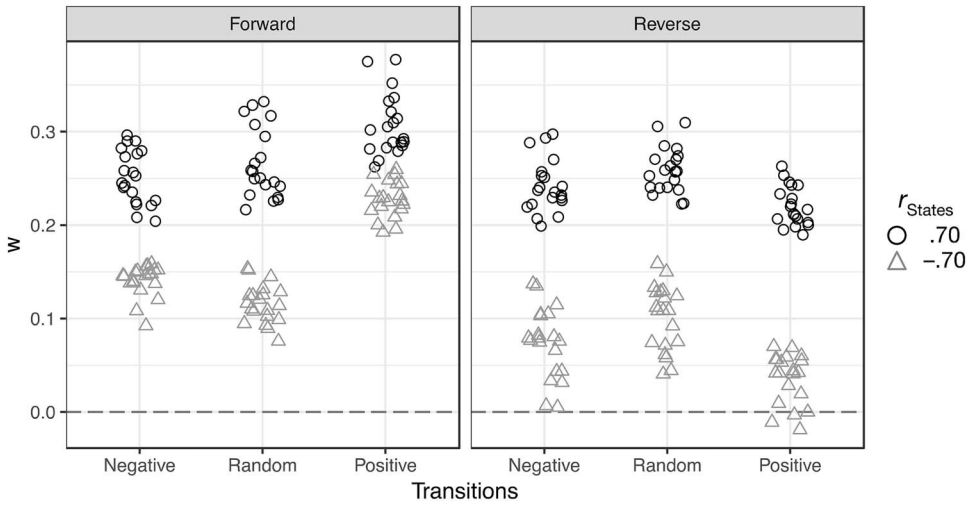

Figure B1. RW Predictions of Causal Strength (*w*) for Stimuli Used in Experiments 1 and 2.

(*Appendices continue*)

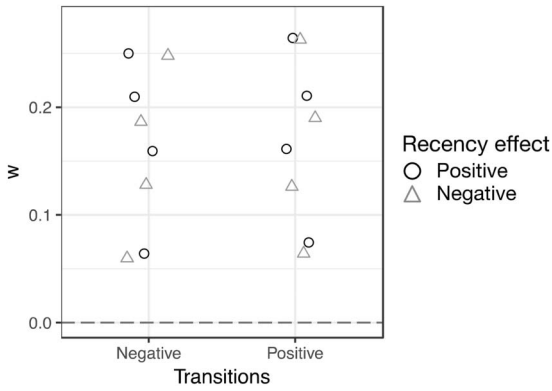

Figure B2. RW Predictions of Causal Strength (*w*) for Stimuli Used in Experiment 3.

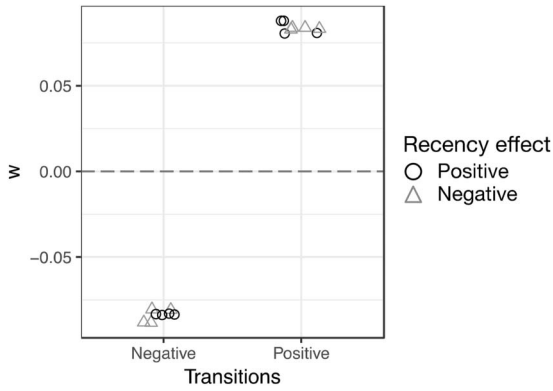

Figure B4. RWΔ Predictions of Causal Strength (*w*) for Stimuli Used in Experiment 3.

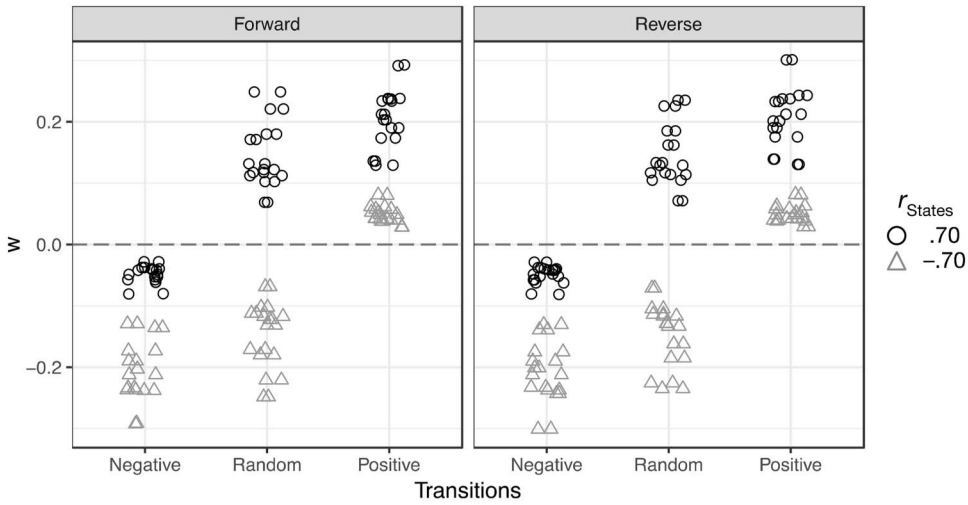

Figure B3. RWΔ Predictions of Causal Strength (*w*) for Stimuli Used in Experiments 1 and 2.

Received February 1, 2017
Revision received February 8, 2018
Accepted February 9, 2018 ■