Elemental Causal Learning From Transitions

by

Kevin Wen Xin Soo

BA in Psychology, HELP University, Malaysia, 2009

MS in Cognitive and Decision Sciences, University College London, 2011

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts & Sciences in partial fulfillment

of the requirements for the degree of

Master of Science in Cognitive Psychology

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS & SCIENCES

This thesis was presented

by

Kevin Soo

It was defended on

December 2, 2015

and approved by

Timothy Nokes-Malach, PhD Associate Professor, Department of Psychology

James Woodward, Professor, PhD, Department of History and Philosophy of Science

Thesis Director: Benjamin Rottman, PhD, Assistant Professor, Department of Psychology

ELEMENTAL CAUSAL LEARNING FROM TRANSITIONS

Kevin Soo, MS

University of Pittsburgh, 2016

Much research on elemental causal learning has focused on how causal strength is learned from the states of variables. In longitudinal contexts, the way a cause and effect change over time can be informative of the underlying causal relationship. We propose a framework for inferring the causal strength from different observed transitions, and compare the predictions to existing models of causal induction. According to this framework, transitions where the cause and effect change simultaneously are the most informative about the underlying causal strength. The predictions of this framework are tested in an experiment where subjects observe a cause and effect over time, updating their judgments of causal strength after observing different transitions. The results are largely consistent with the proposed framework, showing that causal learning in longitudinal contexts relies on patterns of transitions – a previously overlooked source of information from which causal strength can be learned.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Causal knowledge of the world allows us to identify potential causes and explanations for events (e.g. Lagnado, Gerstenberg, & Zultan, 2013; Lombrozo & Gwynne, 2014; Rips & Edwards, 2013) and make decisions about what events to intervene upon to achieve desired outcomes (e.g. Hagmayer & Meder, 2013; Hagmayer & Sloman, 2009; Meder, Gerstenberg, Hagmayer, & Waldmann, 2010). Causal systems are often large and complex (e.g. Hagmayer, Meder, Osman, Mangold, & Lagnado, 2010; White, 1997, 2014) which demand immense computational resources for learning and reasoning about them. However, people get around this by focusing on constrained parts of the whole system (Bramley, Lagnado, & Speekenbrink, 2015; Johnson & Keil, 2014; Lagnado, Fenton, & Neil, 2012). The most basic building block of causal systems is a singular cause-effect relationship, which is the focus of the present research.

The process of learning whether a single potential cause has an influence on an effect (and the strength of that influence) is called elemental causal learning or induction – e.g. Does premium gas get my car better mileage? Does doing a colleague a favor make them more friendly? Does watering my plant twice a week make it healthier than watering it once a week? One way we learn about causal relations is from experiencing these variables over time. The following extended example shows how observations over time can elucidate the nature of the causal relationship between variables.

A patient suffering from chronic fatigue tries a new drug that claims to boost energy levels in an attempt to infer its causal strength. Over the next week, she takes the drug on three days (drug = 1, no drug = 0), keeping track of whether she has high (1) or low (0) energy. Figure 1 shows two possible patterns of experience with the drug, which we contrast.

	Day	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*	*Sun*	*Mon*
(a)	Drug	0	0	1	0	1	0	1
	Energy	1	0	1	0	1	0	0
(b)	Drug	0	0	0	0	1	1	1
	Energy	1	0	0	0	1	1	0

Figure 1. Example Longitudinal Data Sets

From the experience in Figure 1a, how might the patient infer the drug's causal strength? Consider what happens from Wednesday to Sunday. Between each of these days, stopping and starting the drug is accompanied by corresponding changes in energy, suggesting a strong positive causal relationship between the drug and energy. There are other days (Tuesday and Monday) that do not fit with this pattern, so the drug does not always work and it is not always needed. However, given the pattern from Wednesday to Sunday, it is hard to explain the consistent pattern without inferring that the drug had a positive causal influence on energy. It seems less likely that the drug has no influence on energy, but by some coincidence some unknown factors changed the patient's energy at the same times (and in the same direction) that the patient happened to take the drug.

Contrast Figure 1a with Figure 1b. In Figure 1b, the three days that the patient took the medicine are all grouped together. In contrast to Figure 1a, there is a greater possibility that the pattern is due to a coincidence; perhaps the patient had more energy on Saturday and Sunday because it is the weekend, or because she just got over a cold, or her kids are away at camp, etc. Relative to the observations in Figure 1a, it is less convincing that the medicine works in this case. There are practically unlimited numbers of possible alternative causes, and when the trials are grouped together as in Figure 1b it is more plausible that the pattern is merely a coincidence. This example illustrates how the transitions, i.e. the *change* in the cause and effect from one observation to the next convey meaningful information for learning causal strength (e.g., Rottman & Keil, 2011, 2012; Soo & Rottman, 2014, 2015).

1.1 LEARNING FROM STATES VS. TRANSITIONS

Instead of reasoning about transitions, an alternative strategy to learn whether the medicine works is to keep track of the distribution of experienced states. Table 1 summarizes the states in the data from Figure 1; Figure 1a and 1b both contain the same 7 states, observed in a different order. The states are labeled [A], [B], [C] and [D] and their frequencies when summarized in a contingency table are denoted by *a, b, c,* and *d,* respectively.

In Table 1, there are more [A] or [D] states (the cause and effect are often present and absent together) relative to [B] or [C] states (where either the cause or effect is present, but not together), suggesting a positive contingency between the cause and effect (i.e. a positive causal relationship). Many models of elemental causal induction compute causal strength from state frequencies like those displayed in Table 1 (Cheng, 1997; Griffiths & Tenenbaum, 2005; Jenkins

& Ward, 1965; for a review of 41 such computational models of causal induction, see Hattori & Oaksford, 2007). Causal strength is usually computed as a value between -1 (a perfect negative causal relationship) and 1 (a perfect positive causal relationship), with 0 meaning there is no causal relationship.

Table 1. *Frequencies of states in data from Figure 1.*

	Energy = 1	Energy = 0
Drug = 1	*a* = 2	*b* = 1
Drug = 0	*c* = 1	*d* = 3

Note. The drug is the cause and energy is the effect.

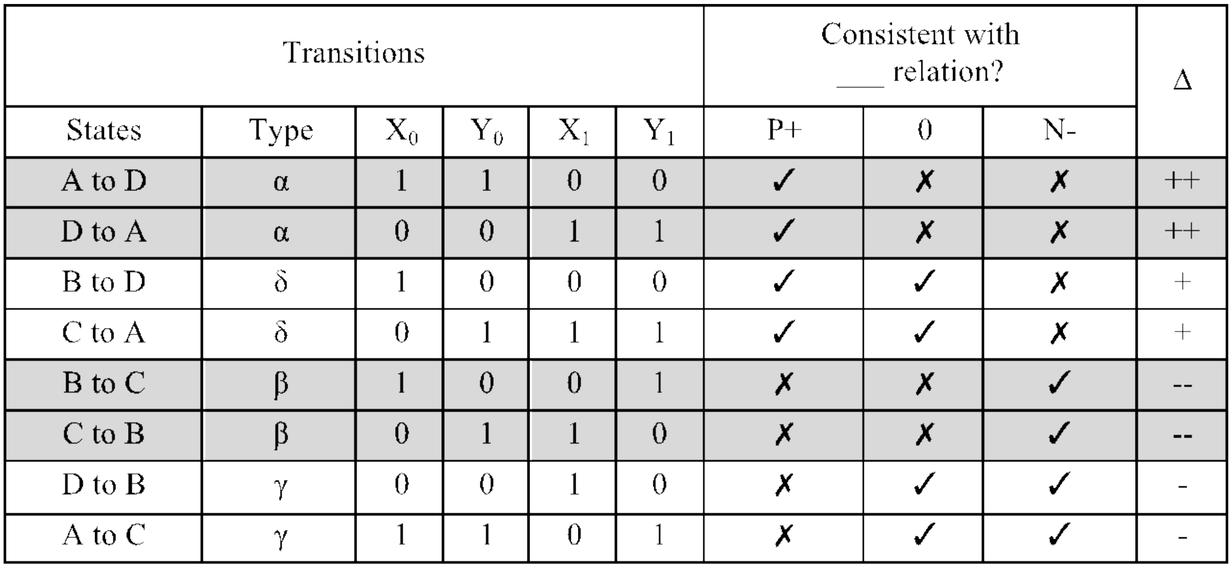
These models are intended for "cross-sectional" situations where each observation is independent of the prior one – e.g. observing 7 patients, where three have taken the medicine and four have not. In cross-sectional situations, the order of the observations is arbitrary – if presented sequentially, the transitions between observations do not convey meaningful information. Because of this, the observations in these contexts could be appropriately presented in summarized form (like in Table 1). In contrast, observations in "longitudinal" contexts have a meaningful order because the variables are unfolding in a single entity across time (Figure 1). In these cases, presenting observations in summarized form would result in crucial information being lost: the data in Figures 1a and 1b can lead to different inferences and explanations when observed in their respective orders, but they would both be summarized into the same table.

Most past studies on causal learning have used cross-sectional cover stories where observations are meant to be interpreted as being temporally independent (though for exceptions, see Hagmayer et al., 2010; Rottman & Keil, 2012; White, 2015). In contrast, real-world causal learning often involves tracking one entity over time due to the sequential nature of experience. In the current study, we are interested in causal learning in longitudinal situations (e.g., tracking one person over time), with the aim of addressing the lacuna in the existing research. In the following section, we propose a framework how different types of transitions could influence learners' beliefs about causal strength based on how they interpret each transition. In the next section, we compare the predictions of this framework to those from existing models of causal induction. Finally, we present behavioral data from an experiment designed to test our framework, showing people are sensitive to transitions over and above states, in patterns generally consistent with the transition-based learning framework proposed here.

1.2 TRANSITION-BASED LEARNING

With a binary cause and effect there are four possible states at any given time point. Thus, there are 4 × 4 = 16 possible transitions that can occur between two adjacent time points in a time series. The framework we propose concerns how each observed transition influences a person's belief about causal strength. Since this research is concerned with learning in a longitudinal context where observations are made sequentially, the framework makes trial-by-trial predictions about how a learner should update her beliefs about causal strength after viewing a particular transition (reflected in judgments of causal strength made after each trial).

Table 2. *Predictions of Transition-Based Learning (TBL) for α, δ, β, and γ transitions.*



Transitions						Consistent with ___ relation?			Δ
States	Type	X_0	Y_0	X_1	Y_1	P+	0	N-	
A to D	α	1	1	0	0	✓	✗	✗	++
D to A	α	0	0	1	1	✓	✗	✗	++
B to D	δ	1	0	0	0	✓	✓	✗	+
C to A	δ	0	1	1	1	✓	✓	✗	+
B to C	β	1	0	0	1	✗	✗	✓	--
C to B	β	0	1	1	0	✗	✗	✓	--
D to B	γ	0	0	1	0	✗	✓	✓	-
A to C	γ	1	1	0	1	✗	✓	✓	-

Note. X_0 and X_1 represents the state of X at time points 0 and 1 respectively. Each transition is shown to be either consistent (✓) or inconsistent (□) with a positive, negative or no (0) causal relation. Δ is the predicted change in causal strength judgment due to the transition. ++ and -- are large changes to causal strength in the positive vs. negative directions, whereas + and – are smaller changes. 0 represents no change to causal strength.

1.2.1 α, δ, β, and γ transitions

Table 2 categorizes 8 of these transitions into 4 types depending on how consistent they are with a positive causal relationship, a negative one, or no relationship. A transition is consistent with a causal relationship if the transition is likely to be generated by a causal relationship of that strength. In general, with positive causal relationships, changes in the cause (X) are accompanied by changes in the effect (Y) in the same direction (e.g. α transitions). With negative causal relationships, changes in X lead to changes in Y in the opposite direction (e.g. β transitions). If there is no relationship, changes in X are not associated with changes in Y. From this logic, one can reason backwards to consider how observing a particular transition should influence one's belief concerning the causal relation.

Consider α transitions – increases in X accompanied by increases in Y ([D to A] transitions), or decreases in X accompanied by decreases in Y ([A to D] transitions). These are transitions that would be generated if X is assumed to have a positive causal influence on Y. Such transitions are unlikely if there is a negative or no causal relationship – one would need to posit a coincidental hidden cause that influenced Y at the same time that X changed (Figure 1a). Since such transitions are most consistent with a positive relation (but not neutral or negative ones), this framework predicts large positive increases in causal strength judgments after α transitions (see first two rows of Table 2).

Next, consider δ transitions such as [C to A] – X increases (0 to 1) but Y stays at 1. This transition is consistent with a positive causal relationship with a ceiling effect for Y (Y cannot increase any higher than 1). A [B to D] transition could be interpreted in the same way but with a floor effect for Y. However, these transitions are also consistent with there being no causal relationship: assuming there is no causal relationship, if X changes, then Y will stay at whatever state is was initially at, all other things being constant. Because α transitions are only consistent with a positive causal relationship while δ transitions are also consistent with no relation, observing α should lead to a larger increase in causal strength judgments than observing δ (though both should lead to an increase). TBL predicts a relative difference in magnitude, but is agnostic about the absolute magnitude of increase for each type of transition.

This logic can be extended to transitions consistent with negative causal relationships. In these cases, an increase in the cause leads to a decrease in the effect, because the cause inhibits the effect. β transitions (only consistent with a negative relation) should lead to larger *decreases* than γ transitions (consistent with a negative relation with a floor/ceiling effect, and also with no relation), in a way analogous to the difference between α and δ transitions.

1.2.2 ε and ζ transitions

ε transitions are when only the effect (Y) changes, while the cause (X) stays the same. With ζ transitions, neither X nor Y change. How these transitions influence one's belief about causal strength depend on the interpretation one assumes of what it means for X to cause Y. Saying that X causes Y can mean multiple things – there is diversity in the labels and language people use to describe or refer to causation. These hint at the diversity of senses that a 'cause' can operate (Levin & Hovav, 1994; Walsh & Sloman, 2011; Wolff & Song, 2003; Wolff, 2007). For example, the statements "X causes Y", "X enables Y", and "X prevents Y" all suggest the state of X having some sort of causal influence over the state of Y.

The following sections consider two different interpretations of what it can mean for X to cause Y in a longitudinal context, leading to a branching of our TBL framework into two versions (TBL-PUSH and TBL-PUSH-HOLD). The two versions make diverging predictions for ε and ζ transitions (see Table 3a and 3b). These different interpretations of X causing Y do not concern the transitions described above (α, δ, β, and γ).

1.2.3 TBL-PUSH: cause as ‘push’ only

In the transitions described above, X is assumed to cause Y in the sense that a change in X causes a change in Y (unless Y is already at floor or ceiling). We call this the ‘push’ interpretation – the change in X is transmitted to Y through the causal link. Causal verbs like ‘force’, ‘set’, ‘stimulate’ and ‘start’ are associated with this sense of causation (Wolff & Song, 2003; Wolff, 2007).

Consider the interpretation of ε transitions (only Y changes) when causation is understood in this manner. If causation means that a change in X ‘pushes’ Y, then when X stays the same during ε transitions, X is not acting causally upon Y. Changes in Y alone can be explained by the influence of some unobserved variable – i.e. Y was ‘pushed’ by an unobserved cause. There can be a positive, negative or no relation between X and Y, and Y can still change independently of X. This is because while X remains stable, it is not transmitting anything causal to Y. Therefore, ε transitions are uninformative of the causal strength of X.

Next, consider ζ transitions, where neither X nor Y change. With a ‘push’ interpretation of cause, these transitions are consistent with there being no causal relation between X and Y: X and Y simply remain in the same state as before. These transitions are also consistent with there being a positive or a negative causal relation: X does not change and is not ‘pushing’ Y, so Y is not expected to change. Because they are consistent with all possible causal relations, these transitions are also uninformative of X’s causal strength.

Because these transitions are uninformative, the ‘push’ causation version of TBL predicts no change in causal strength judgments upon observing these transitions. These predictions are presented in Table 3a, showing the TBL-PUSH branch of our framework.

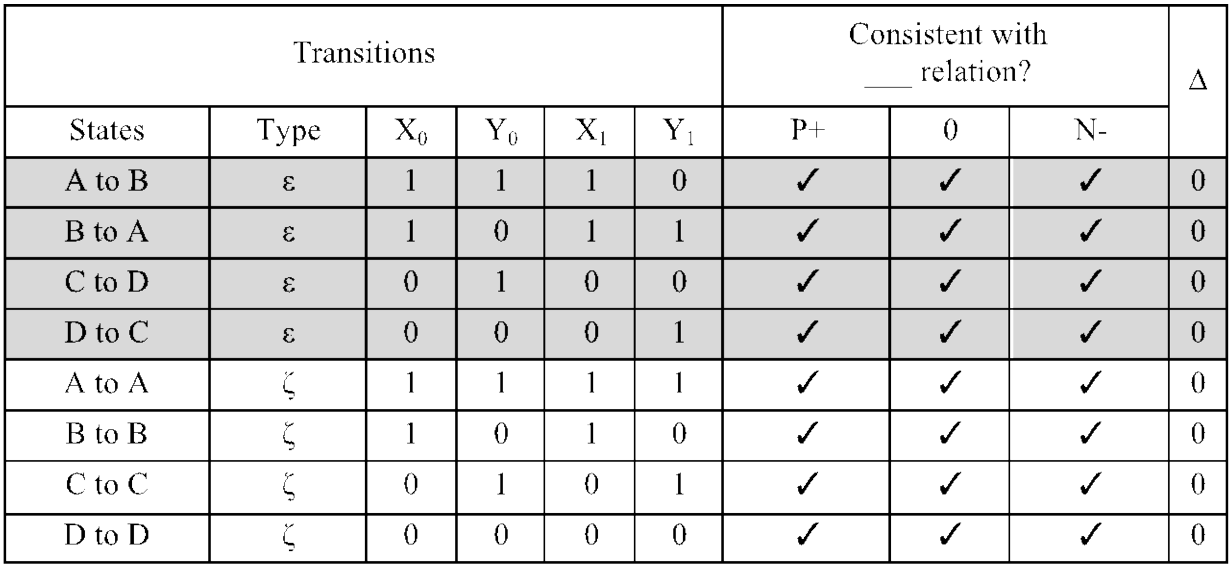
1.2.4 TBL-PUSH-HOLD: cause as ‘push’ and ‘hold’

TBL-PUSH assumes a ‘push’ interpretation of causation. However, in a longitudinal context, it also makes sense to talk about the causal powers of X operating continuously over time on Y, referred to here as the ‘hold’ interpretation of causation. In addition to ‘hold’, this sense of causation is conveyed through causal verbs like ‘keep’, ‘protect’ and ‘restrain’ (Wolff & Song, 2003). When X holds Y, it means that X being in a particular state is what causes Y to stay in or return to a particular state (e.g. if X = 1 and there is a positive relation, Y = 1 as well). This interpretation of causation assumes that X’s causal strength acts continuously on Y, rather than being transmitted to Y only when X’s state changes (the ‘push’ interpretation). TBL-PUSH-HOLD incorporates both the ‘push’ and ‘hold’ conceptions of causation.

Under TBL-PUSH-HOLD, ε transitions can imply either a positive or negative causal relation as well. [A to B] and [D to C] transitions are where X and Y start out in the same states at the start of the transition, and then Y changes. These can be interpreted as X ‘failing to hold’ Y in the same state, which is evidence against a strongly positive relation and for either no relation or a negative relation. TBL-PUSH-HOLD therefore predicts a decrease in causal strength judgments for both of these transitions. The converse logic applies for [B to A] and [C to D] transitions – X ‘fails to hold’ Y in the opposite state, implying the causal relation is not as strongly negative as would have been believed at the beginning of that transition. TBL-PUSH-HOLD predicts an increase in causal strength judgments for these transitions.

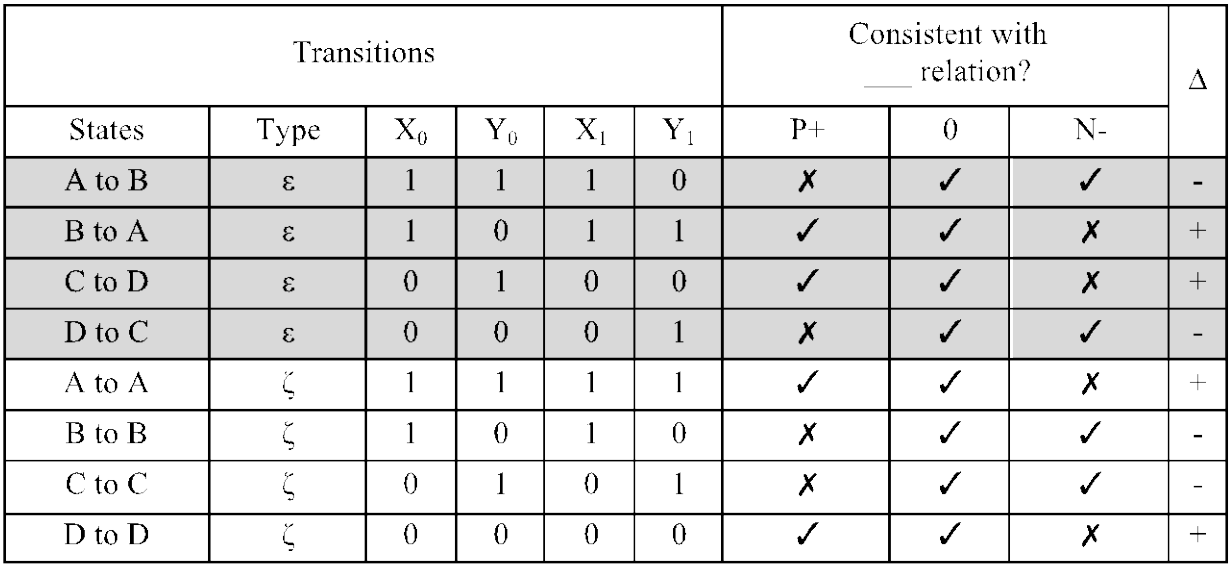
Next, consider the predictions for ζ transitions under TBL-PUSH-HOLD. [A to A] and [D to D] transitions can be interpreted as X ‘successfully holding’ Y in the same state, which is consistent with a positive causal relation or no causal relation, but not with a negative causal relation. TBL-PUSH-HOLD predicts that causal strength judgments should increase on these transitions because they are evidence of positive strength ‘hold’ causation. The converse is true for [B to B] and [C to C] transitions, where observations provide evidence of ‘hold’ causation with a negative relation, or there being no causal relation.

Table 3. (a) *Predictions of TBL-PUSH for ε and ζ transitions with only ‘push’ causation.*



Transitions						Consistent with ___ relation?			Δ
States	Type	X_0	Y_0	X_1	Y_1	P+	0	N-	
A to B	ε	1	1	1	0	✓	✓	✓	0
B to A	ε	1	0	1	1	✓	✓	✓	0
C to D	ε	0	1	0	0	✓	✓	✓	0
D to C	ε	0	0	0	1	✓	✓	✓	0
A to A	ζ	1	1	1	1	✓	✓	✓	0
B to B	ζ	1	0	1	0	✓	✓	✓	0
C to C	ζ	0	1	0	1	✓	✓	✓	0
D to D	ζ	0	0	0	0	✓	✓	✓	0

(b) *Predictions of TBL-PUSH-HOLD for ε and ζ transitions with ‘push’ and ‘hold’ causation.*



Transitions						Consistent with ___ relation?			Δ
States	Type	X_0	Y_0	X_1	Y_1	P+	0	N-	
A to B	ε	1	1	1	0	✗	✓	✓	-
B to A	ε	1	0	1	1	✓	✓	✗	+
C to D	ε	0	1	0	0	✓	✓	✗	+
D to C	ε	0	0	0	1	✗	✓	✓	-
A to A	ζ	1	1	1	1	✓	✓	✗	+
B to B	ζ	1	0	1	0	✗	✓	✓	-
C to C	ζ	0	1	0	1	✗	✓	✓	-
D to D	ζ	0	0	0	0	✓	✓	✗	+

1.2.5 Summary of predictions

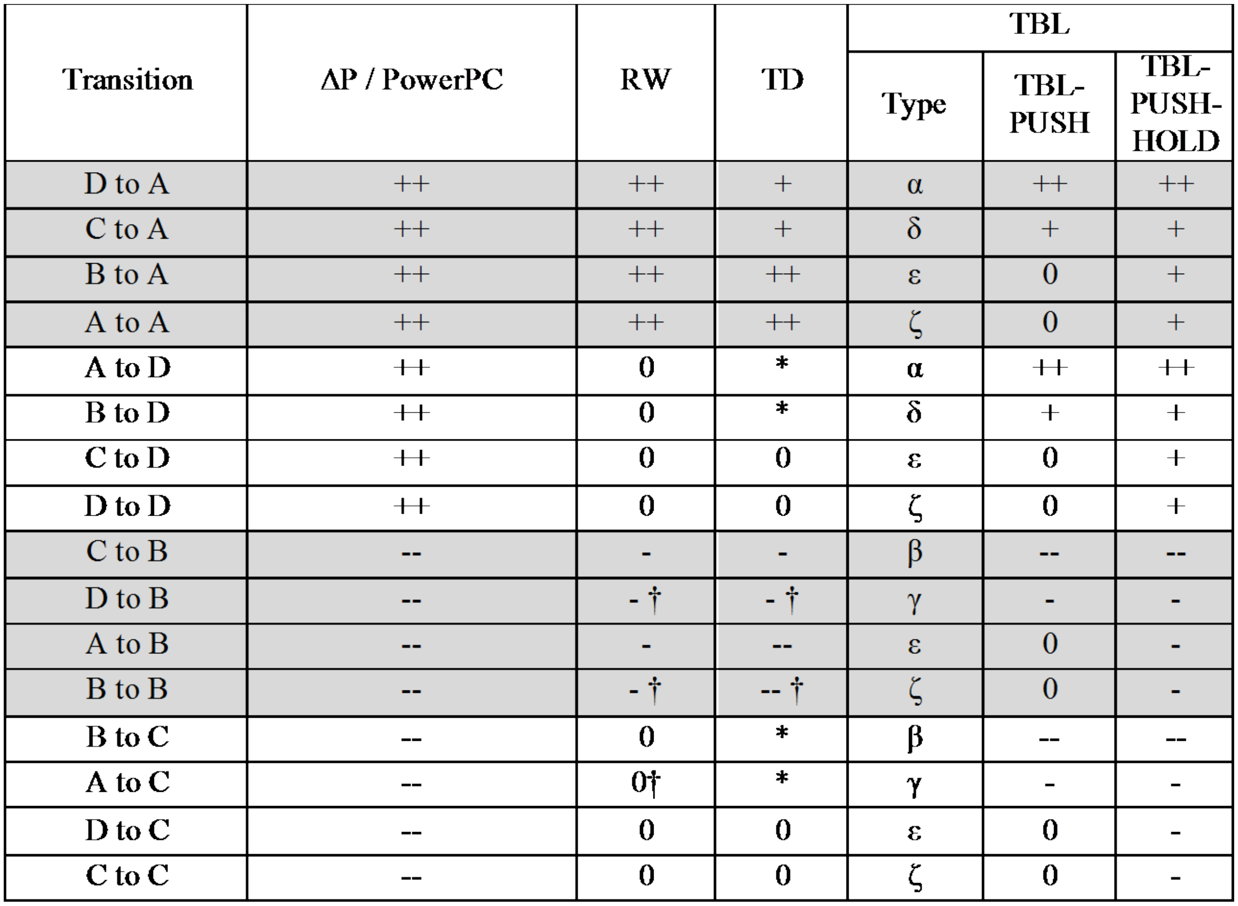
The TBL framework makes predictions for α, δ, β, and γ transitions in Table 2. The framework then makes two competing sets of predictions for ε and ζ transitions depending on whether one includes a conception of 'hold' causation (Tables 3a and 3b). In sum, the most crucial prediction made by this theory is that α and β transitions will lead to more change (in the positive and negative direction respectively) than all other types of transitions that lead to changes in causal strength judgments. From here, both versions are referred to collectively as TBL. If one specific version is referenced, the names TBL-PUSH and TBL-PUSH-HOLD will be used.

1.3 MODELS OF CAUSAL INDUCTION

In Table 4, TBL's predictions are compared with several existing models of causal strength learning. In the following sections in the introduction, these models are presented at a high-level theoretical perspective, to show all but one of the models (TD) are not inherently sensitive to transitions the way they are defined in the previous sections. However, there are a variety of complex reasons that various models can appear to predict different amounts of change in causal strength for different transitions. These complex explanations are discussed thoroughly in Appendix A, for readers who are particularly interested in those models.

Because the other models are mainly sensitive to states, not transitions, Table 4 groups together the four transitions that end in the same state (in shades of gray vs. white). This highlights the role of transitions for TBL above and beyond states. Within each of the four groups of transitions in Table 4, the first row are α or β transitions (both variables change), the second are δ or γ (ceiling and floor effects), the third are ε transitions (effect changes by itself), and the fourth are ζ transitions (neither variable changes). This is the organization that will be maintained for purposes of discussing the various model predictions, and also for analyzing the results of the experiment.

Table 4. *Model predictions for 16 transitions.*



Transition	ΔP / PowerPC	RW	TD	TBL		
				Type	TBL-PUSH	TBL-PUSH-HOLD
D to A	++	++	+	α	++	++
C to A	++	++	+	δ	+	+
B to A	++	++	++	ε	0	+
A to A	++	++	++	ζ	0	+
A to D	++	**0**	*****	**α**	++	++
B to D	++	**0**	*****	**δ**	+	+
C to D	++	**0**	**0**	**ε**	**0**	+
D to D	++	**0**	**0**	**ζ**	**0**	+
C to B	--	-	-	β	--	--
D to B	--	- †	- †	γ	-	-
A to B	--	-	--	ε	0	-
B to B	--	- †	-- †	ζ	0	-
B to C	--	**0**	*****	**β**	--	--
A to C	--	**0†**	*****	**γ**	-	-
D to C	--	**0**	**0**	**ε**	**0**	-
C to C	--	**0**	**0**	**ζ**	**0**	-

Note. The predicted changes for ΔP/PowerPC assume that the prior data is held constant, and that the causal strength is not already at ceiling or floor. ++ and -- denote a predicted increase or decrease that is larger relative to + and – within the same model. 0 denotes no predicted change. * These cases depend upon too many factors so no generalized predictions can be made. † These cases lead to no change on the rare occasions when the effect has not been present in the data on an earlier trial.

1.3.1 ΔP and Power-PC

ΔP (Jenkins & Ward, 1965) and Power PC (Cheng, 1997) are two of the most prominent normative models of causal strength induction, and both calculate causal strength from the frequencies of each state in the contingency table (e.g. Table 1). They both produce a causal strength rating ranging from -1 to 1. ΔP is calculated using equation (1) and Power-PC is calculated using equations (2) and (3):

$$\Delta P = \frac{a}{a+b} - \frac{c}{c+d} \qquad (1)$$

$$PowPC\ (positive\ causal\ strength) = \frac{\Delta P}{1 - {}^{c}/_{c+d}} \qquad (2)$$

$$PowPC\ (negative\ causal\ strength) = -\frac{\Delta P}{{}^{c}/_{c+d}} \qquad (3)$$

ΔP computes causal strength as the difference in the probability of the effect when the cause is present vs. when it is absent. Cheng’s (1997) Power-PC is a more theory-driven model of causal strength, which accounts for the possibility of other unobserved causes that could also cause the effect. It calculates causal strength as the relative increase or decrease in the probability of the effect due to the target cause, rather than the absolute difference.

In the present study, we are investigating the effects of transitions using trial-by-trial judgments, so ΔP and Power-PC will be computed after each trial. For both of these models, after an [A] or [D] observation, the causal strength judgment will go up unless the causal strength is already at ceiling (1), and after a [B] or [C] observation the judgment will go down unless the causal strength is at floor (-1).

One way to show how these models are not sensitive to transitions involves considering four sequences of data all ending in A: [A, B, C, D, A], [D, A, B, C, A], [C, D, A, B, A], and [B, C, D, A, A]. In all four of these sequences the causal strength ratings predicted by ΔP and Power-PC would be exactly zero after the 4th trial because all these sequences contain one each of [A], [B], [C] and [D] at that point. Then, after the 5th trial, the causal strength rating would increase. However, it would increase by exactly the same amount under all four sequences. After the fifth trial, causal strength would be 1/6 for ΔP and 1/3 for Power-PC. In contrast, TBL makes different predictions for the change in the causal strength judgment after the fourth vs. after the firth trial, because the sequences of data have different transitions. The predictions for ΔP and Power-PC, which are insensitive to transitions, are shown in the second column in Table 4.

1.3.2 RW (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972)

The Rescorla-Wagner model is a model of associative learning that has also been proposed of human causal learning (Shanks & Dickinson, 1987; Danks, 2003). RW is a trial-by-trial model of learning that updates weights representing the strength of the association between a cue (cause) and outcome (effect) after each observation. RW was created to model change in associative strength within a single animal over time, so unlike the models mentioned above it is meant to handle longitudinal data. RW computes ΔV_X, the change in associative strength for the cause (X) after a given trial. In the case of a single cue (the single cause, X), this is computed with equation (4) below:

$$\Delta V_X = \alpha\beta(\lambda_Y - V_{Total}) \qquad (4)$$

α and β are parameters representing the salience of the cue and the context in which the outcome appears, respectively. They are treated here as a single learning rate parameter. λ represents the intensity of the outcome or effect. In the case of a binary effect (Y), this will take on either a value of 1 or 0. V_{Total} is the aggregate associative strength of all cues present on that trial. In the case of elementary causal learning, this includes the existing strength of X and the background, which is always present.

RW was developed to model many temporal phenomena in how associative strengths get updated over time like acquisition curves and blocking. Thus, RW is sensitive to the order of the data unlike the state-based models above. RW works by updating the associative strength after each observation. At any trial, the change in the associative strength is calculated based on the difference in the error prediction of the outcome (effect) from summing the associative strengths of the present cues (in this case, X and the background). Thus, the factors determining how associative strength is updated are the newly observed state, and the existing associative strength - the state observed in the immediately prior trial is irrelevant.

The current associative strength constrains the magnitude that the associative strength can change when a new observation is experienced. If the current associative strength is zero (e.g. on the first trial before any reinforcement happens) and an [A] trial occurs, there will be a relatively large increase to the associative strength due to the large error in the prediction (because $\lambda_{\mathrm{Y}} = 1$ and $\Delta_{\mathrm{Total}} = 0$). In contrast, if the current associative strength is 0.75 and an [A] trial occurs, there will be a smaller increase in the associative weight (because the error is smaller).

Provided that the associative strength is not already at asymptote (+1 or -1), the strength will always increase on an [A] trial, with the magnitude of change determined by the existing associative strength of the cause and the background. To ensure that the associative strength stays within the bounds, a low learning rate parameter is usually chosen.

Changes to the strength only occur when the cue is present, so no changes occur after [C] and [D] trials; though Van Hamme & Wasserman (1994) have proposed that the strengths be updated even when the cue is absent.

The strength will always decrease on a [B] trial, with one exception. The causal strength cannot decrease on a [B] trial if the effect has never been present previously in the data – there has never been an [A] or [C] trial. If an [A] or [C] trial has never occurred, then the associative strength is zero. Then, when a [B] trial occurs (cause present, effect absent), there is zero prediction error, so the associative strength for the cause is zero.

In general, [A] trials lead to changes (increases) of larger magnitudes [B] trials do. This is due to some of the inhibitory strength associated with the cause being offset by the positive associative strength of the context. This difference is reflected in Table 4 – RW predicts large increases (++) for transitions ending in [A], and small decreases (–) for transitions ending in [B].

In sum, even though RW is sensitive to many aspects of the order of the observations, the specific prior observation does not have any impact on the change in the associative weight; all that matters is the newly experienced observation, the learning rate parameters, and the prior associative strength. In other words, RW is insensitive to transitions. This means that RW calculates the same amount of change after an [A] trial regardless of the prior state, accounting for the current associative weight. This is reflected in Table 4 – RW makes the same predictions for all of the transitions in each group ending in the same state.

1.3.3 TD (Sutton & Barto, 1987)

Temporal difference (TD) learning is a form of reinforcement learning (also see Gallistel & Gibbon, 2000; Sutton, 1988). Here, we discuss a particular instantiation of TD learning that can be used to model classical conditioning paradigms. Even though TD is heavily based on RW and it has been widely applied in other areas of psychology (cf. Seymour et al., 2004), as far as we know TD has never been proposed or analyzed as a model of human causal learning. The weight (*w*) of a particular cue (in this case, the cause, X) at the next time point, *t* + 1, represents the change in the learner's belief about causal strength when exposed to new data. It is computed using equation (5):

$$w_{t+1} = w_t + c[r_t + \gamma P(w_t, x_t) - P(w_t, x_{t-1})]\bar{x}_t \qquad (5)$$

r_t is the presence or absence of the effect at time point *t*, *c* is a scaling parameter. $P(w_t, x_t)$ is a prediction function that calculates the presence of the effect, computed from whether the cue (*x*) is currently present, and its weight (w). The parameter γ is the *discount rate* (discussed below).

The important features of TD will be discussed by contrasting it with RW. First, compared to RW, learning (or changes to the associative strength) occurs repeatedly moment-to-moment as opposed to just once at the end of the trial. Our simulations break each trial into 10 "moments" in time.

Second, whereas RW learns weights that minimize the prediction of the unconditioned stimulus (effect) at a given trial, TD predicts a sum of future values of the effect signal discounted such that the near future is weighted more than the distant future. This can be seen in the error term, found in the square brackets of equation (5). The term $r_t + \gamma P(w_t, x_t)$ represents the causal strength at time point *t* (in TD, given by w_t) computed from a weighted sum of causes, $P(w_t, x_t)$, when information about the effect at that time point is known (adding r_t to the term). The error in TD is this quantity's difference from the prediction of w_t from information at the prior time point, before information about the effect (r_t) is known. Hence, $P(w_t, x_{t-1})$ is subtracted from $\gamma P(w_t, x_t)$. In other words, TD models associative strength based on the difference between predictions of the effect at successive time points. TD tries to predict the associative strength of the cause at *t* from the information up to $t – 1$, and then looks at the difference that information at *t* makes to the prediction (knowing the effect, r_t). Reinforcement occurs when new information at *t* results in a large difference to the prediction (i.e. is 'surprising' to the prediction from $t – 1$).

The parameter γ is called the *discount rate* that determines the rate that values of *r* (state of the effect) at the new time point (*t*) get discounted by weighting the prediction from the cause. Values closer to the upper bound of 1 will result in the onset and offset of a cause generating more reinforcement than its mere presence. An earlier version of the TD model had this built in by default – the onset and offset of the cause was reinforcing (Sutton, 1988).

Third, whereas RW only updates weights for the cause when the cause is present, TD updates weights for the cause in proportion to the strength of an "eligibility" trace, *x*-bar, which is equivalent to a memory/salience trace for the cause. *x* increases to asymptote as long as the cause is present, and then decays when it is absent, both in ways approximating a power function. If the cause has been present for a while (e.g. observing transitions that start with [A]), its level of "eligibility" increases. This then scales the amount by which the prediction error can reinforce the weight of the cue, as seen by how the error term is multiplied by *x*-bar. On the other hand, if the cause was recently absent and *x*-bar is low, learning is slow until the eligibility trace increases. Even after the cause disappears some learning can occur to the extent that the trace persists (e.g. in [A to C] transitions).

The final difference is that unlike ΔP, Power-PC and RW, the weights for the cues (*w*) in TD are not bounded between -1 and 1.

All these features and dynamics in TD mean that unlike the other models discussed so far, TD actually makes predictions based on transitions, not just states. However, TD is quite complex and the exact predictions are sensitive to its multiple parameters. One important determinant of TD's updating behavior is the trace. In order to simplify the predictions of TD to present them in Table 4, we focus on qualitative predictions that are determined from the trace. Appendix A presents details of simulations from best-fitting parameters. As can be seen in Table 4, even though the trace makes TD sensitive to certain transitions, the transition-based predictions of TD are very different from TBL.

First, TD predicts a greater increase for [A to A] and [B to A] than [C to A] or [D to A] transitions. In the latter two transitions the trace of the cause (*x*-bar) is initially zero, and it takes time to increase, meaning reinforcement occurs at a slower rate. In the former two transitions, the cause is already present from the previous time point, so *x*-bar is initially higher, speeding up learning at the present time point. The difference between [B to B] and [A to B] vs. [C to B] and [D to B] is also due to the eligibility trace, but in the negative direction. However, just like RW, on the occasions when the effect has not been present in the data prior to a transition ending in [B] (e.g. if [D to B] or [B to B] transitions occur first), then TD predicts there is no change because there has not yet been any effect to be predicted.

In the transitions [C to C], [C to D], [D to C], and [D to D], the cause is not present so the trace is zero, and thus its weight is not updated. This is similar to how RW does not update strength when the cue is absent. The transitions [A to D], [B to D], [B to C] and [A to C] are extremely dynamic and depend on the prior weight, w_t (e.g. whether it is > 0 or < 0), the prior weight of the unobserved cue (the background) and the parameter values in TD. Thus, we cannot make a generalized characterization of how the weights get updated for these transitions.

1.3.4 Comparisons between models

Predictions for the transition-based learning (TBL) framework (both TBL-PUSH and TBL-PUSH-HOLD versions) are included in the two rightmost columns in Table 4 (note the different ordering of the transitions from Tables 2 and 3). There is some consistency in all of the models. Transitions ending in [A] and [D] are viewed as positive (or neutral) evidence for all models, whereas those ending in [B] and [C] are negative (or neutral) evidence for all models.

There are several comparisons that will help distinguish between the models. Firstly, ΔP, Power PC and RW make the same predictions for all transitions ending in the same state, unlike TBL, which makes different predictions for transitions within a group that end in the same state. Secondly, though both TD and TBL are sensitive to transitions, the predictions are in certain cases highly divergent, especially for the transitions ending in A. Thirdly, according to both versions of TBL, the α and β transitions are the most distinctive in that they predict the largest increases (or decreases), and this distinction is not present in any of the other models. For this reason, the analyses in the experiment will focus on the differences between α and β compared to other transitions ending in the same state. Lastly, TBL-PUSH and TBL-PUSH-HOLD can be distinguished by examining whether ζ and ε transitions lead to equivalent or smaller changes as δ and γ transitions.

The goal of our experiment was to investigate whether different models of causal induction explain different aspects of longitudinal causal learning. The comparisons outlined above, will be informative in contrasting the different models.

2.0 EXPERIMENT

Subjects observed sets of longitudinal data and made causal strength judgments after each trial. We were focused on whether the changes in subjects' causal strength judgments on a given trial were influenced by the transition they had just observed. The experiment reported here was the last of five similar versions that were conducted before. A description of the previous versions and a discussion of the problems associated with them necessitating various iterations can be found in Appendix B.

2.1 METHOD

2.1.1 Subjects

100 subjects were recruited through Amazon Mechanical Turk (MTurk). The experiment was conducted online and took roughly 15-20 minutes to complete. An additional 5 subjects dropped out without completing the full experiment, but we included their data in the final analysis. Of this 105, 18 subjects were excluded from the final analysis because their pattern of responses suggested a misinterpretation of the task and response scale used in the experiment (discussed below in the results section).

2.1.2 Design and stimuli

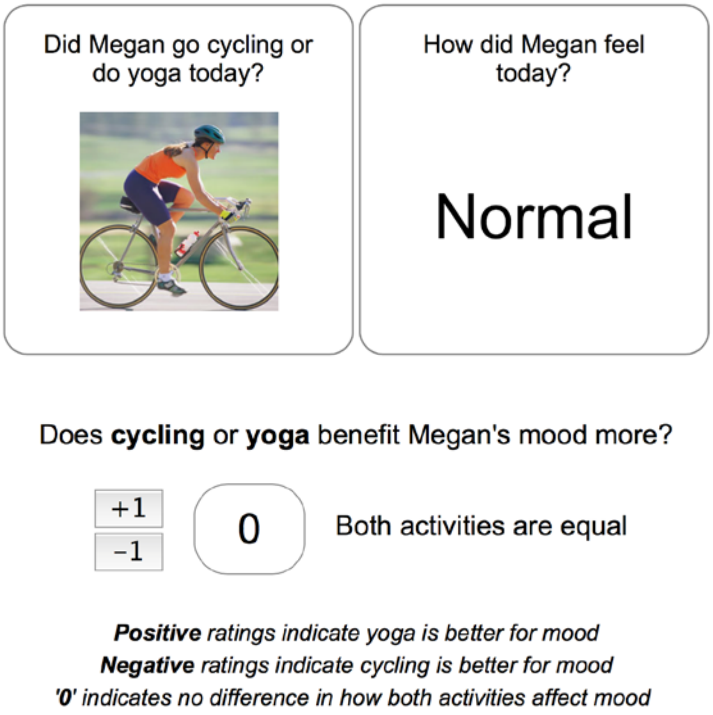
Subjects were presented with sets of data consisting of a binary cause and effect. Each subject viewed 7 sets of data and one demo scenario beforehand to familiarize them with the task. Each set (one scenario) consisted of 8 trials. The demo scenario consisted of a sequence of alternating [D] and [A] states, resulting in seven α transitions. This was the clearest case (according to our TBL framework) of the strongest possible positive causal relationship, and served as a clear demo that subjects could try after reading the instructions. Additionally, it allowed us to check their use of the response scale.

The 7 scenarios consisted of data sets with differing state distributions, chosen to result in different overall contingencies as computed by Power-PC: 1 / 0.67 / 0.5 / 0 / -0.5 / -0.67 / -1 (see Appendix C for the state distributions corresponding to each Power-PC rating). The trials within a data set were randomly ordered. The reason for having data sets with positive, negative, and neutral contingencies was to have a sampling of all the transitions. For example, β transitions are impossible in the Power-PC = 1 data sets, but common when Power-PC = -1. Each data set had 8 states, resulting in 7 transitions. Subjects experienced the different scenarios in a random order.

2.1.3 Procedure and materials

Subjects were told to imagine they were researchers studying how different morning routines affected people's moods. Each scenario involved observing a volunteer trying out either their regular morning routine or a new routine for a series of 8 days (each 'day' = 1 trial). Examples of routines were doing yoga, cycling, reading a book, sleeping in, etc. At each trial, subjects viewed whether the volunteer did their regular routine or a new routine that morning, and their mood for that day that the volunteers reported at the end of the 'day' (either 'Normal' or 'Happy').

The display subjects viewed consisted of two boxes where the cause and effect were displayed. At the beginning of the scenario, the boxes are empty except for the questions above to indicate what each box displays (see Figure 2). When the subject clicks on a button to progress, a picture appears in the left box indicating which of the routines the volunteer did that day and text appears in the right box indicating their mood that day (as seen in Figure 2). After viewing the data for a particular day, subjects estimated the causal strength of the new routine on the volunteer's mood using a scale ranging from -99 to 99 that appeared beneath the stimuli.



Did Megan go cycling or do yoga today?

How did Megan feel today?

Normal

Does **cycling** or **yoga** benefit Megan's mood more?

Both activities are equal

Positive *ratings indicate yoga is better for mood*
Negative *ratings indicate cycling is better for mood*
'0' *indicates no difference in how both activities affect mood*

Figure 2. Experimental stimuli shown to subjects and response scale.

Negative ratings indicate that the regular routine is better for mood than the new routine, and positive ratings indicate that the new routine is better. The magnitude (distance from 0) indicates the strength of that causal relationship. A rating of 0 indicates that there is no difference between the two routines in their influence on the volunteer’s mood. Subjects clicked on buttons to increase or decrease their estimates, adjusting them from the value displayed on the screen.

After doing so, subjects clicked on a button to progress to the next trial. The picture (indicating the routine) and text (indicating mood) in the boxes fade out and the picture and text for the next trial appear (this transition takes about one second). Subjects then adjusted their estimate, and then progressed to the next trial. They repeated this until they viewed data for all 8 days. On each new trial (after the initial one), their causal strength estimate from the previous trial remained on display and was the starting point from which they adjusted their estimate.

Although the cover story and scale seem to require that subjects judge the strength of two different causes (the old and new routines), having them judge the difference in their effects on the volunteer’s mood is similar to judging a single cause (the new routine). Here, the new activity (1) is contrasted against the old activity (0), whereas more traditional formulations of binary causal relationships show the cause as being either present (1) or absent (0). Prior iterations of the experiment suggest that presenting the stimuli in this new way helped with subjects’ understanding of the response scale (see Appendix B).

Subjects viewed 7 scenarios, each one involving a different volunteer trying out different pairs of activities. There were no repeats of any of the activities.

2.2 RESULTS

18 subjects were excluded from the analysis because their responses indicated a misinterpretation of the scale: on [D] states; they tended to reduce their causal strength judgments. Further investigation revealed that these subjects' judgments tracked the occurrence of the effect; they generally increased judgments when the effect was present (even on [C] trials), and decreased them when it was absent. This interpretation occurred in this subset of subjects despite efforts at defining a positive vs. negative relation (see method section). We did not want to further train subjects on the use of the scale by providing feedback. This was to avoid implying that there was one right answer and interpretation of the stimuli, and to preserve subjects' natural use of the scale as much as possible. However, this interpretation of the scale is not explainable by any of the models of causal induction (see Table 4). Thus, we excluded those who made decreases on more than 50% of [A to D] transitions they experienced in the actual scenarios (there was some variation in how many of these transitions each subject experienced because of how the stimuli were generated).

Data for 87 subjects was included in the final analysis. 82 completed the 7 scenarios, and the remaining 5 completed a total of 13. This resulted in data for 587 scenarios (4109 transitions). The variable under analysis was the change in causal strength judgments (the difference between the judgment at the present and prior trials) for each of the 16 types of transitions. This was analyzed with four separate regression models, one for each group of transitions ending in the same state (shaded groups in Table 4). The α and β transitions (top row within each of the 4 sets in Table 4) were the reference transition, as the most important hypothesis was whether α and β transitions produced larger changes than other transition types.

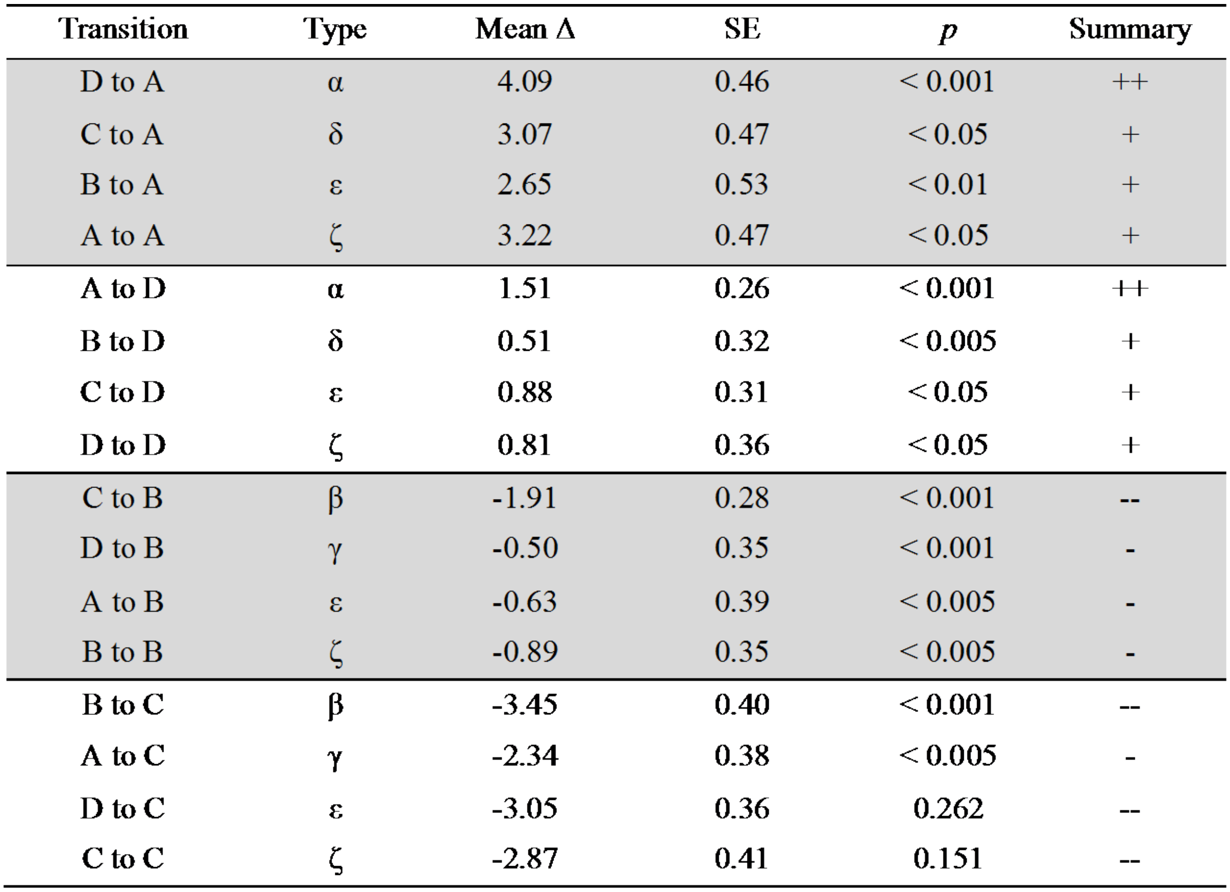
Each of the regression models controlled for the causal strength at the prior trial, as it was possible it could constrain the amount of change a subject made. For example, for transitions expected to lead to an increase, a very high starting point could constrain the amount a subject could possibly increase their judgment by. The mean change in causal strength judgments reported here are after controlling for prior strength. However, in examining the distribution of responses, subjects generally made small increases or decreases and used only the middle regions of the scale. Thus, ceiling and floor effects are not a major factor – there were only seven transitions in the entire data set where subjects were at ceiling or floor. A random intercept for each subject was included in each regression, as each subject made multiple judgments.

Table 5 displays the results of the four regressions (indicated by the different shades). Within the four transitions ending in a given state, the other three transitions were all compared against the top transition (α and β). In typical regression tables, the difference between levels is reported. But for ease of interpretation the differences are translated here to their own group means (e.g. [C to A] transitions produced an average increase of 3.07 points). The rightmost column in Table 5 summarizes how the transitions within that regression differ from each other. All transitions yielded average change scores that were significantly different from 0.

At first glance, the most obvious result is that all transitions ending in [A] and [D] produced increases in causal strength, whereas those ending in [B] and [C] produced decreases. Considering only the direction of change, this finding is most consistent with ΔP and Power-PC. However, there are differences in the magnitude of change within each group of transitions, and the relative differences are somewhat consistent with the predictions of TBL.

The most critical result is that α and β transitions always produced the largest changes within each group of transitions. In the case of transitions ending in [A], [B] and [D], α and β transitions led to larger changes than all other transitions within the same group. This finding is uniquely predicted by TBL (both versions); it is not predicted by the other models and is even the opposite prediction made by TD. In the case of transitions ending in [C], the [B to C] transitions (β) did not lead to significantly larger decreases than for [D to C] and [C to C] transitions (ε and ζ), but the differences in means were in the predicted direction. All these transitions produced larger decreases than [A to C] (γ) transitions. The pattern for this group of transitions was partially consistent with state-based models (β being equal with ε and ζ) and partially with TBL (β leading to larger decreases than γ).

Table 5. *Regression results for effect of transitions. Separate models for each group of transitions.*



Transition	Type	Mean Δ	SE	p	Summary
D to A	α	4.09	0.46	< 0.001	++
C to A	δ	3.07	0.47	< 0.05	+
B to A	ε	2.65	0.53	< 0.01	+
A to A	ζ	3.22	0.47	< 0.05	+
A to D	α	1.51	0.26	< 0.001	++
B to D	δ	0.51	0.32	< 0.005	+
C to D	ε	0.88	0.31	< 0.05	+
D to D	ζ	0.81	0.36	< 0.05	+
C to B	β	-1.91	0.28	< 0.001	--
D to B	γ	-0.50	0.35	< 0.001	-
A to B	ε	-0.63	0.39	< 0.005	-
B to B	ζ	-0.89	0.35	< 0.005	-
B to C	β	-3.45	0.40	< 0.001	--
A to C	γ	-2.34	0.38	< 0.005	-
D to C	ε	-3.05	0.36	0.262	--
C to C	ζ	-2.87	0.41	0.151	--

Note. The Summary column indicates the relative increase or decrease of a particular transition relative to the other transitions in the same group. Thus, ++ simply refers to a significantly greater increase than +, which in turn refers to a significantly greater increase than 0. They do not refer to the absolute magnitude of change relative to other groups. *p*-values of α and β rows are for a test of these means vs. 0, *p*-values of other rows are for tests of those means relative to the baseline condition of that regression.

The findings related to α and β transitions do not support RW, which predicts the same magnitude of increase for all transitions ending in [A] and the same magnitude of decrease for all transitions ending in [B]. In addition, RW predicts no change for transitions ending in [C] and [D], but subjects do make changes on these transitions. The findings were also not consistent with TD. Within the groups of transitions ending in [A] and [B], TD predicts the largest changes for ε and ζ transitions. The findings were opposite to this – for transitions ending in [A] and [B], subjects made larger increases for the α/β transitions. The differences between transitions ending in the same state are also inconsistent with the state-based models, which predict no transition effects.

In the groups of transitions ending in [A], [B] and [D], ε and ζ transitions led to changes that were smaller than α and β transitions, as predicted by TBL. In these groups, the change was equivalent to that of the δ and γ transitions, consistent with TBL-PUSH-HOLD rather than TBL-PUSH. Looking at the relative magnitude of changes associated with each transition, the results are most consistent with TBL-PUSH-HOLD.

2.2.1 Model fits

Fitting the models to the participants' data was accomplished in the following way. ΔP and Power-PC do not have any parameters, so no fitting was required. One unique feature of these models is that these models cannot be calculated until at least one [A] or [B] trial has been observed and at least one [C] or [D] trial has been observed. Trials that do not meet these criteria were removed from the analysis.

Fitting RW required first choosing an appropriate learning rate parameter. In equation (4) above, it can be seen that the parameters in RW are a product and can be combined into a single learning rate parameter that is greater than 0 (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). We searched for the best fitting parameter values between 0.01 and 0.99 with increments of 0.01, that yielded the highest correlation with subject' trial-by-trial changes in causal strength ratings. The highest correlation (r = 0.34) was found using the lowest learning rate parameter, so the parameter value of 0.01 was chosen. A low learning rate parameter means that learning occurs slowly, and changes in the causal strengths are very small (Table A1). However, when assessing the models, we focus on the relative magnitudes of the predicted changes in causal strength, within a model, rather than the absolute predicted changes in causal strength across models, so the small changes are not problematic.

Fitting TD required a number of steps. Because TD is a real-time model rather than one based on trials – it expects as input, not a single trial with a particular observation like [A], but continuous input that is sampled at a certain rate (say, 10 times a second). Consequently, each trial was elongated into 10 time points. The change in the causal strength for a given trial is the difference of the weight at t_{10} vs. t_1.

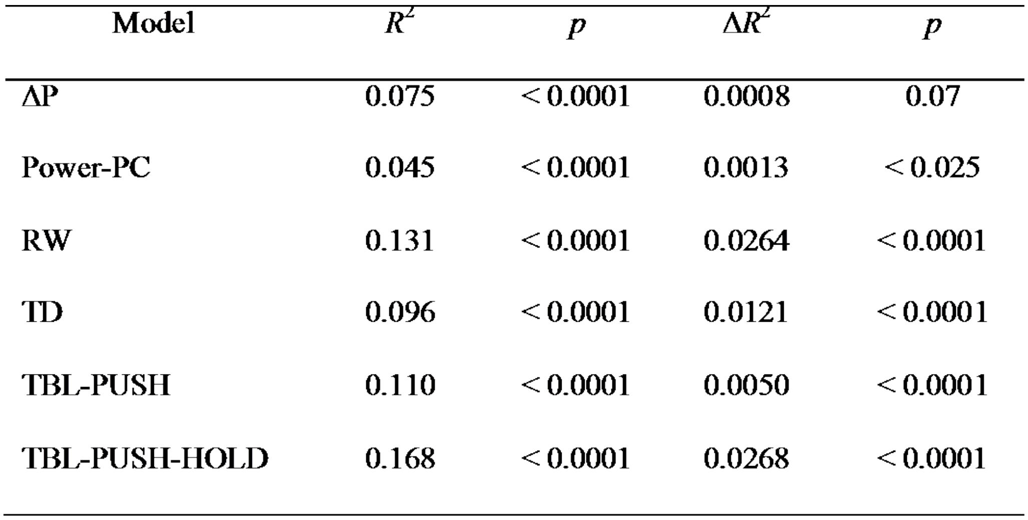
Like with RW, TD was fitted with parameters that would lead to the highest correlation between the model's predicted change and subjects' actual change on each trial. Three parameters, c, γ and β, were fit with a grid search on every combination of parameters between 0.1 to 1 in increments of 0.1. The best combination was $c = 0.1$, $\gamma = 0.9$ and $\beta = 0.1$, which resulted in a correlation of $r = 0.31$ between TD's predictions and subjects' changes on each trial. $\gamma = 0.9$ is close to 1, which suggests the data is best fit when TD's reinforcement is influenced more by the onset and offset of the cause (suggesting transitions are more important relative to states). $\beta = 0.1$ is small, which means the eligibility trace of the cause grows slowly when the cause appears and decays quickly after the cause disappears. The scaling parameter c does not have an intuitive interpretation in the original formulation by Sutton and Barto (1987). (TD also had a couple other combinations of parameters that resulted in local maxima. We chose to report this combination because it resulted in the best predictive validity of subjects' responses and the parameters were interpretable.)

Each of these models was used to predict the change in subjects' causal strength ratings for each transition across all the data they experienced. A series of regression models were used to predict subjects' responses from the model predictions. In this analysis, we did not control for the prior strength judgment because the various models do this. Each of these regressions sought to answer the following questions. First: does a particular model account for a significant amount of variance in subjects' responses? Second: does a particular model account for a significant amount of variance in subjects' responses over and above all the other models?

The second goal was to see how well the predictions of TBL fit subjects' responses compared to the other models. The TBL framework makes predictions about the relative magnitudes of change for each transition, which can be interpreted as an ordinal scale. Thus, to match the predictions in Table 4, the predictions were coded as follows: ‘++’ was coded as 2, ‘+’ as 1, ‘–’ as -1, ‘– –’ as -2, and no change as 0 (note the different prediction patterns for TBL-PUSH vs. TBL-PUSH-HOLD, as shown in Table 3).

Table 6 presents the fits of the various models. The R^2 column represents how much variance a particular model accounts for. As can be seen, each of the models explains a small but significant amount of variance of subjects’ responses. Of all the models, TBL-PUSH-HOLD accounts for the most variance in subjects’ responses.

Table 6. *Fits for each model to subjects’ responses.*



Model	R^2	p	ΔR^2	p
ΔP	0.075	< 0.0001	0.0008	0.07
Power-PC	0.045	< 0.0001	0.0013	< 0.025
RW	0.131	< 0.0001	0.0264	< 0.0001
TD	0.096	< 0.0001	0.0121	< 0.0001
TBL-PUSH	0.110	< 0.0001	0.0050	< 0.0001
TBL-PUSH-HOLD	0.168	< 0.0001	0.0268	< 0.0001

Note. R^2 represents the amount of variance in subjects’ responses a particular model accounts for on its own. ΔR^2 represents the unique variance in subjects’ responses a particular model accounts for after controlling for all the other models.

The second question of interest is whether each model (in particular, both versions of TBL) explains unique variance that is not accounted for by other models. The ΔR^2 column represents the added variance accounted for by a particular model over and above all other models. This was computed by running a regression predicting subjects' responses using all models and comparing it to a regression using all models except for the model of interest, giving us the difference in R^2. As can be seen, each of the models explains a significant amount of unique variance, including both versions of TBL. TBL-PUSH-HOLD accounts for the most unique variance of all the models.

3.0 GENERAL DISCUSSION

Previous work studying elemental causal learning has focused on how causal strength is learned from states – [A], [B], [C] and [D]. In the present manuscript, an extreme version of an elemental causal learning theory that focuses exclusively on transitions was proposed. This proposal was intended to be provocative – to theorize how different transitions could be interpreted completely independently of states. In reality, it is unlikely that people exclusively rely upon transitions at the exclusion of states. Indeed, the results suggest that both state-based and transition-based models independently predict subjects' causal strength judgments. The main finding in support of TBL was that when both the cause (X) and effect (Y) changed (α and β transitions), participants changed their causal strength judgments more than when only the cause changed (δ and γ), when the same state was repeated (ζ), and when the effect changed on its own (ε), controlling for the prior causal strength judgment. This is a unique prediction of TBL that is not made by any of the other models.

Two different versions of TBL were proposed. In TBL-PUSH, X is said to cause Y when a change in X produces a change in Y – called the "push" conception of causation. In TBL-PUSH-HOLD, "push" causation is assumed, but in addition, X is said to have a causal influence on Y when it "holds" Y in the same state when X stays in the same state over time. The findings that ε and ζ transitions produced significant changes to the causal strength supported the TBL-PUSH-HOLD model over TBL-PUSH.

As mentioned above, the transition-based predictions of TBL, especially TBL-PUSH were deliberately extreme, while in reality it is possible that humans use a combination of state and transition-based information. The pattern of results consistent with TBL-PUSH-HOLD over TBL-PUSH is in line with this view. While an extreme transition-based view would emphasize only "push" causation, cross-sectional data (when observations are temporally independent) assumes only a "hold" conception of causation because causation must be judged entirely from the states of the X-Y pairs (there is no meaningful change in the variables between pairs, and therefore no "push" causation that can be detected). In longitudinal contexts, causation can theoretically involve both "push" and "hold" conceptions – X can cause Y by both transmitting a change and by transmitting its strength continuously to keep it in a particular state. For example, taking a drug can affect a biological system both when it is introduced ("push") and by being maintained in the system ("hold"). Because TBL-PUSH-HOLD includes both "push" and "hold" conceptions of causation, it is the model that theoretically comes closest to the idea that people use both state and transition-based information.

3.1 AMBIGUOUS INTERPRETATIONS OF TRANSITIONS

As hinted at above, there are multiple ways that ζ transitions, when neither the cause nor effect changes, could be interpreted. ζ transitions could be interpreted either as two consecutive observations of the same piece of data, or as just one observation. The interpretation would likely be moderated by factors that the reasoner uses to segment time. This issue also arises when comparing continuous-time models like TD vs. trial-based models like RW. Some research has investigated learning in continuous time (Buehner & May, 2009; Pacer & Griffiths, 2012).

Beuhner & May (2009) found that when observing a cause and effect in a continuous event stream, the lack of clearly delineated trials combined with delays between cause and effect makes it difficult for subjects to attribute an effect's presence to a particular cause.

In the present research we tried to avoid this ambiguity by partitioning observations into discrete trials. This was necessary to isolate and test the effects of transitions as well as to situate this research within the standard trial-by-trial causal learning paradigm. However, in truly continuous-time learning settings, it is not clear how learners will handle periods of time in which neither the cause nor effect change (ζ transitions). In continuous-time situations when there are no trials to discretize time, perhaps such periods will not lead to much updating, in line with TBL-PUSH-HOLD.

ε transitions, when the cause stays the same but the effect changes, can also be interpreted in multiple ways. In the cover story used in the experiment here it was most plausible that the cause (an exercise routine) would have an influence on the effect (the volunteer's mood) on the same trial (same day). Assuming that these causes had effects on the same day, ε transitions where only Y changes should be attributed to some unobserved factor influencing Y. However, if a learner permits longer influences of the cause, other interpretations could arise. Consider the sequence X = [0, 1, 1] and Y= [0, 0, 1]. A learner could attribute the change in Y at Time 3 to the change in X at Time 2.

From the discussion above, it is evident that transitions may be interpreted in multiple ways. There is some interesting research on how observed states can be interpreted differently given different prior knowledge – in some instances even [A] can be interpreted as negative evidence (Luhmann & Ahn, 2011). For instance, if subjects believe that the causal strength is actually inhibitory, observing the cause and effect being present can suggest either that the cause fails to inhibit the effect (standard interpretation), or that some unobserved factor is responsible for the fact that the effect is present, but that the cause is still inhibitory. When reasoning about states, beliefs about unobserved factors drive the different interpretations. We also hypothesize that beliefs about unobserved factors are responsible for some of the interpretations of transitions, particularly for the difference between TBL-PUSH vs. TBL-PUSH-HOLD.

Future work should delve deeper into the myriad of possible interpretations for each transition, perhaps via methods like eliciting verbal explanations. Additionally, manipulating beliefs about how unobserved causes operate in a causal learning scenario should be able to elicit different interpretations of particular transitions.

3.2 BEYOND BINARY VARIABLES

The majority of research on causal learning has focused on binary causes and effects, even though human reasoners often learn about variables that take on states with much finer gradients than simply being present or absent. Much causal inference in the social sciences (e.g. economics, political science) involves tracking variables that are measured on a continuous, or at least ordinal scale.

On the one hand, there are theories of causal reasoning such as Power-PC (Cheng, 1997) for which the presence/absence of binary variables plays a critical metaphysical role. On the other hand, other research suggests that people dichotomize continuous variables into binary variables for tractability (e.g. Marsh & Ahn, 2009). Though there has been some research on causal reasoning with continuous/ordinal variables (Hagmayer et al., 2010; Pacer & Griffiths, 2011; Soo & Rottman, 2014; White, 2015), the vast majority has focused on binary variables.

There are several ways of how continuous variables might affect the causal induction process and how the TBL framework would need to be reformulated to accommodate continuous variables. Firstly, a characteristic of binary variables is that there is a very strict ceiling and floor – which is the reason our present formulation of TBL needs to incorporate δ and γ transitions (where a change in the cause is not accompanied by a change in the effect, because it might already be at ceiling/floor). With ordinal or continuous variables, there is a much greater range for the states a variable can take on, and ceiling or floor effects would be much less frequent (or impossible, depending on the domain and variables in question). This might make causal induction using the TBL framework easier, as most transitions involving a strong causal relationship would involve both X and Y changing together, with Y rarely (if ever) exceeding the ceiling or floor – this would result in transitions that would have been counted as δ and γ in the binary case becoming α and β transitions, which are much more informative for causal induction in longitudinal contexts.

Causal learning with continuous variables would also change the interpretation of ζ transitions – with a sensitive enough measurement scale; a variable should never stay exactly the same, meaning that ζ transitions are theoretically impossible. However, it is likely that a variable can 'seem' to remain stable, either due to a coarse measurement scale (e.g. the value of X and Y are rounded to integers) or due to human reasoners disregarding small variations as unimportant or uninformative about the causal process. For example, Soo & Rottman (2014) demonstrated that human reasoners could distinguish changes in a variable that were due to the influence of a potential cause from noise (small variations that were uninformative).

3.3 CONCLUSIONS

The current results suggest that elemental causal learning in longitudinal contexts involves a combination of transition and state-based reasoning. One important goal for future research is to better capture how these two types of reasoning get used – do they get used simultaneously, are there individual differences, or does a single learner sometimes use one strategy and other times use another?

Future research should also investigate the factors that can promote the use of one reasoning pattern over another, possibly by varying how salient the longitudinal context is to subjects. It will be interesting to see if humans use transition or state-based reasoning in a rational manner suited to the learning and inference goals in a particular environment.

The present research is a first step towards what is hoped will ultimately result in a theory of causal reasoning with longitudinal data.

APPENDIX A

MODEL PREDICTIONS

This appendix provides more details about how each of the models perform, and compares the average change score for each model after each of the 16 transition types observed in the datasets that participants experienced. One finding that will be noticed immediately when comparing the results of this analysis (Table A1) with the predictions in the introduction (Table 4), is that the average amounts of change in the causal strength vary by transition even for transitions ending in the same state (e.g., [D to A] vs. [C to A]), which appears to contradict the claim in the introduction that ΔP, Power-PC, and RW are not sensitive to transitions.

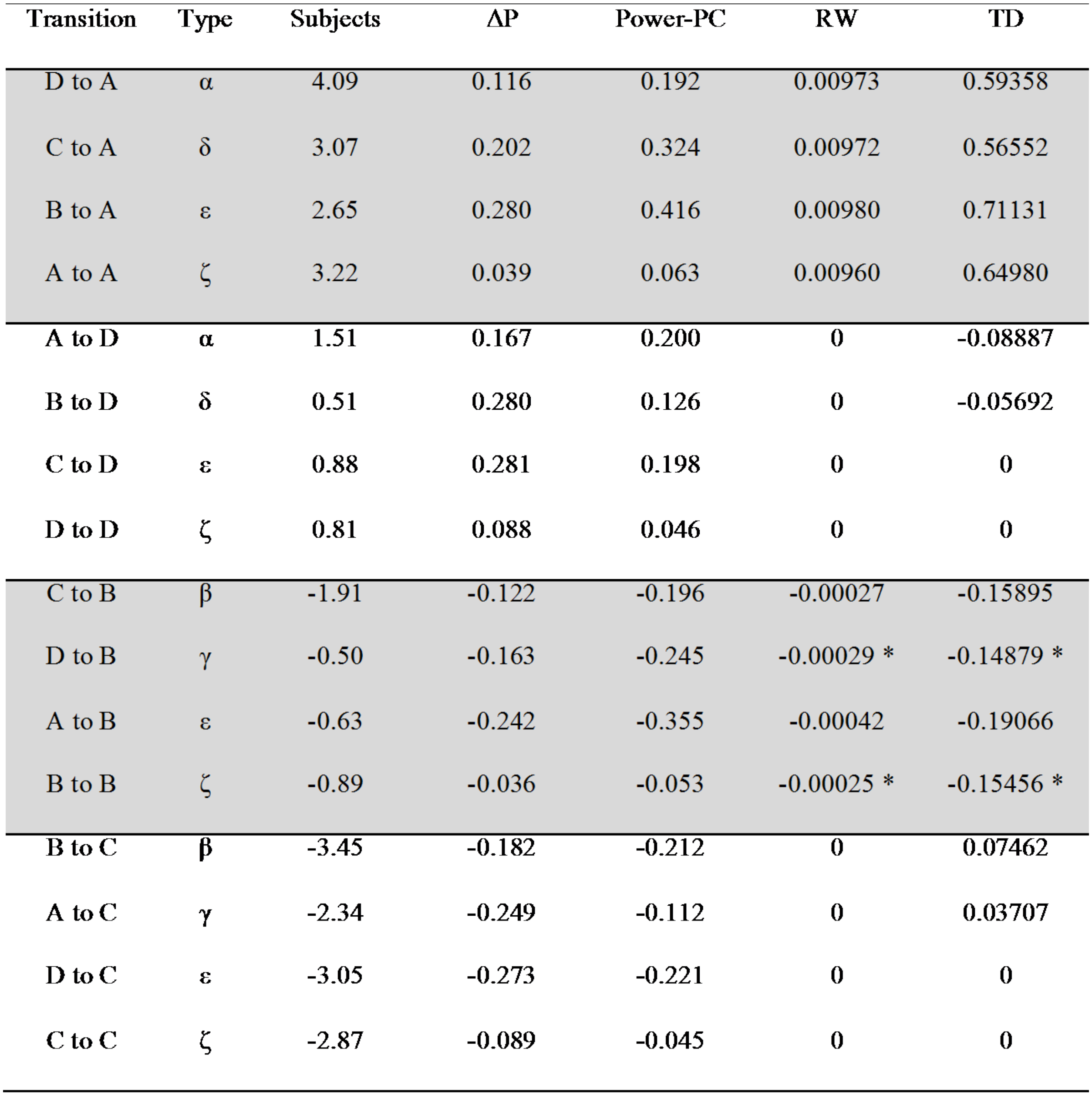
The reason for the apparent contrast is that unlike the examples in the introduction, which hold the prior trials in the dataset constant, when looking at every transition in all the datasets that participants experienced, the prior trials are not held constant. The choice to conduct the experiment with randomly ordered data, as opposed to creating highly controlled comparisons like those given in the introduction (e.g., [A, B, C, D, A] vs. [D, A, B, C, A]) was made for two reasons. First, allowing the order of the trials to vary randomly, and analyzing all transitions in the dataset allows for more external validity than a very narrow set of comparisons. Second, collecting data on all 16 transition types would require a separate condition for each transition type, and would only permit collecting small amounts of data for a given comparison (e.g., [D to A] vs. [C to A]). In the current experiments all transitions were analyzed, regardless of the order of the data, allowing for much more data to be analyzed.

The following sections detail the predicted changes in strength for each of the 16 transition types predicted by ΔP, Power PC, RW, and TD, for the set of data that participants experienced. The average change in causal strength computed by each model for each transition in the data is presented in Table A1, along with the mean changes in subjects' responses (controlling for prior strength) for each transition in the same data (from Table 5) for the sake of comparison. Because the models make predictions along differing scales, the predictions for each model will be discussed only by comparing different transitions within a single model, not across models.

Overall, the main finding is that the predictions made by these models deviate systematically from both TBL as well as from the human data.

For RW and TD, a small number of [D to B] and [B to B] transitions did not lead to any predicted decreases because they appeared in data sets without the effect having appeared before them. 56 [D to B] transitions and 58 [B to B] transitions were removed from the present analysis. The average changes for those transitions as computed by RW and TD presented in Table A1 are exclude these 114 transitions. Thus, those means (indicated by * in TableA1) are more precisely read as the average decrease on [D to B] and [B to B] transitions computed by RW and TD when those models do predict a decrease.

Table A1. *Mean change for each transition type made by subjects and simulated with different models.*



Transition	Type	Subjects	ΔP	Power-PC	RW	TD
D to A	α	4.09	0.116	0.192	0.00973	0.59358
C to A	δ	3.07	0.202	0.324	0.00972	0.56552
B to A	ε	2.65	0.280	0.416	0.00980	0.71131
A to A	ζ	3.22	0.039	0.063	0.00960	0.64980
A to D	α	1.51	0.167	0.200	0	-0.08887
B to D	δ	0.51	0.280	0.126	0	-0.05692
C to D	ε	0.88	0.281	0.198	0	0
D to D	ζ	0.81	0.088	0.046	0	0
C to B	β	-1.91	-0.122	-0.196	-0.00027	-0.15895
D to B	γ	-0.50	-0.163	-0.245	-0.00029 *	-0.14879 *
A to B	ε	-0.63	-0.242	-0.355	-0.00042	-0.19066
B to B	ζ	-0.89	-0.036	-0.053	-0.00025 *	-0.15456 *
B to C	β	-3.45	-0.182	-0.212	0	0.07462
A to C	γ	-2.34	-0.249	-0.112	0	0.03707
D to C	ε	-3.05	-0.273	-0.221	0	0
C to C	ζ	-2.87	-0.089	-0.045	0	0

Note. In computing the mean predicted change for each transition, some trials were excluded. For ΔP and Power-PC, there were no predictions made for some observations that appeared in a scenario before there were sufficient observations of all states to compute a prediction.

A.1 ΔP AND POWER-PC

On average, ΔP and Power-PC increase on [A] and [D] trials, and decrease on [B] and [C] trials, as predicted in Table 4. There are some exceptions to this rule. In a sequence like [A, D, A], ΔP and Power-PC are both at ceiling (1) after the [D] trial, and cannot increase with the final [A]. Even a sequence like [A, C, A], which is not at ceiling after the [C] trial, does not result in an increase because the quantity a / (a + b) in Equation 1 does not change.

One feature of the ΔP and Power-PC data in Table B1 that stands out in contrast to Table 4, is that different transitions ending in the same state have different changes in strength, on average, which appears to contradict the claim that ΔP and Power-PC are not sensitive to transitions. In reality, these differences arise because the transitions have different starting trials, and consequently comparing the average change scores for different transitions does not hold the total experience constant.

For example, compare the following two sequences, which display ΔP scores after each trial in superscript: [C^{NA}, D^{NA}, $A^{.5}$, B^{0}, $A^{.17}$] and [A^{NA}, B^{NA}, $C^{-.5}$, D^{0}, $A^{.17}$]. Both sequences produce the same change in ΔP for the final transition, an increase of .17 on the [B to A] and [D to A] transitions. In this example, the prior experience is held constant because after the fourth trial, there is one trial of each type, resulting in a ΔP of 0.

However, in the datasets that participants experienced, the prior trials were not held constant. For example, consider comparing the changes in causal strength during [D to A] transitions vs. [B to A] transitions. Because [D to A] transitions start with a [D] trial, and [B to A] transitions start with a [B] trial, if ΔP is calculated after the [D] trial, it is on average going to be higher than after the [B] trial. This means that [D to A] transitions, on average, occur after more positive evidence, than [B to A] transitions. For example, compare the following two sequences, which display ΔP scores after each trial in superscript: [A^{NA}, C^{0}, $B^{-.5}$, $A^{-.33}$] vs. [A, C^{0}, $D^{.5}$, $A^{.5}$]. In the first sequence, the [B to A] transition starts lower (-.5) and leads to a .17 increase in ΔP, whereas in the [D to A] transition starts higher (.5) leads to zero increase in ΔP. This means that when comparing different transition types within the data that participants experienced, ΔP and Power-PC will appear to show some systematic differences in causal strength for different transitions that end in the same state even though they are not sensitive to transitions.

Most importantly, the average change scores for the 16 transition types for ΔP and Power-PC do a poor job of predicting subjects' responses (Table B1), and they are also inconsistent with the predictions of TBL. Some of the patterns (e.g., [C to A] vs. [D to A]) directly contradict the empirical data.

Additionally, subjects were almost never constrained by the ceiling or floor – subjects' strength judgments in this data set were at ceiling or floor on only seven transitions. Both ΔP and Power-PC, being highly sensitive to ceiling and floor effects, fails to characterize this aspect of subjects' causal strength judgments.

A.2 RW

The best-fitting learning rate parameter led to predicted changes with very small magnitudes (Table 5). However, since our analysis focuses on relative magnitudes of different transition types within a model, rather than absolute magnitudes, this is not a problem. The changes in Table B1 match closely to the theoretical predictions of RW in Table 4, with just one exception; the decreases associated with [A to B] transitions are larger (more negative) than the other transitions ending in [B]. The reason is that during an [A to B] transition, the strength starts higher due to the initial [A] trial. For the other transitions it is possible that no [A] trial had ever been experienced prior to the transition, thus they start lower. Having a higher strength at the start of the transition would result in more error when [B] is observed, leading to a larger decrease. Empirically, there are great differences between the predictions of RW with subjects' responses (in Table B1).

A.3 TD

As can be see from Table 5, the predictions of TD largely corresponded to the predictions we outlined in Table 4. TD predicted increases for all transitions ending in [A], decreases for all transitions ending in [B], and no changes for [C to D], [D to D], [D to C] and [C to C] transitions. In addition, the [B to A] and [A to A] transitions were larger than the [D to A] and [C to A] transitions. The remaining transitions are dependent on particular combinations of the parameters, and thus no generalized predictions could be made in Table 4. Most importantly, the

TD predictions do not qualitatively match the pattern of subjects' responses, and in some cases (e.g., [D to A] vs. [B to A]) directly contradict the patterns in the empirical data.

APPENDIX B

CHANGE IN MATERIALS

The experiment reported above was the last of five versions. As stated above, some subjects' responses revealed a misinterpretation of the scale: their judgments tracked the presence of the effect rather than the strength of the cause. We identified these subjects as those who decreased their judgments of causal strength on more than 50% of [A to D] transitions they encountered. In earlier versions of the experiment, between 30-40% of subjects used the scale in this way (compared to 18% in the final version reported above). In addition to a possible misinterpretation of the scale, perceptual or pragmatic effects could have driven some of these responses, as well.

Here, several key aspects of the experiments that were changed over the course of the numerous iterations to address these issues are described. Each of these reveals interesting psychological processes that are implicated in subjects' interpretation of the experimental stimuli. However, that discussion is beyond the scope of the present study.

B.1 RESPONSE SCALE

Earlier versions made use of cover stories related to testing drugs that had causal effects on the level of various chemicals in the bloodstream. The cause was either present or absent, and the effect was either low or high. A crucial difference between earlier versions and the final version was the response scale used. In earlier versions, subjects used a slider (as shown below) to adjust their causal strength judgments.

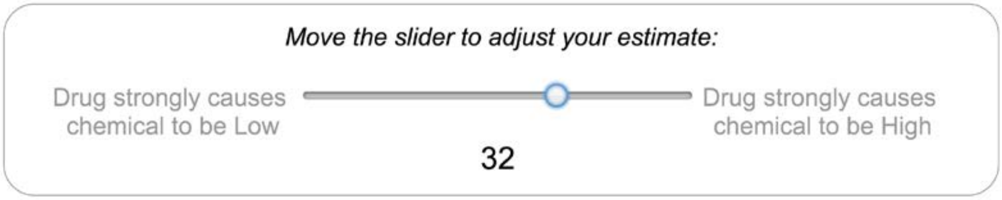

Move the slider to adjust your estimate:
Drug strongly causes chemical to be Low
Drug strongly causes chemical to be High
32

Figure B1. Slider used in earlier iterations of response scale.

This encouraged subjects to make large changes at every transition, and many subjects sometimes used the extreme ends of the scale (switching it from -99 to 99 and vice-versa) depending on the transition, which made it difficult to discern changes of differing magnitudes as predicted by the various models. Additionally, the scale meant it was too easy to make changes in judgment, leading to very noisy data – and in many cases, large decreases on [A to D] transitions.

A decision was made to switch to buttons that increased or decreased judgments one point at a time to discourage this. This was somewhat successful, as indicated by the relatively low average change scores in the reported experiment.

B.2 VISUAL PRESENTATION OF STIMULI

The picture below shows how the stimuli were presented in some early versions. The button on the left representing the state of the cause would slide left if the cause was absent, and right if the cause was present. The label for the effect would also slide left and right when it changed. These animations were included to make the transitions noticeable – if there was an instantaneous change, subjects may not have noticed the transitions. However, it was possible that the animations from left to right mapped visually onto the slider used in earlier versions (see prior section). [A to D] transitions involved both the cause and effect sliding from right to left, which subjects may have mapped on to the response scale and making large changes to the left (decrease in causal strength).

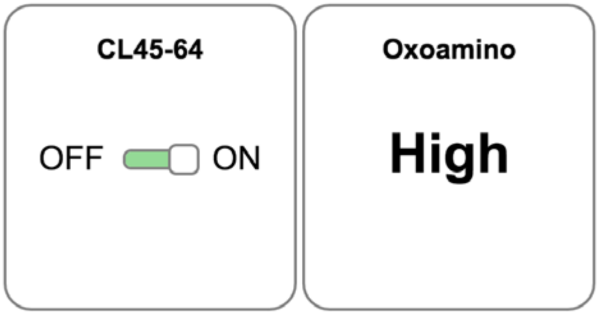


Figure B2. Visual presentation of stimuli used in earlier iterations of the experiment.

To address this, subsequent versions used alternative presentations. We eliminated the animation of stimuli moving left and right, replacing it with a fade-in-fade-out animation to eliminate any mapping of the states of the variables onto the direction of movement.

B.3 STIMULI FORMAT

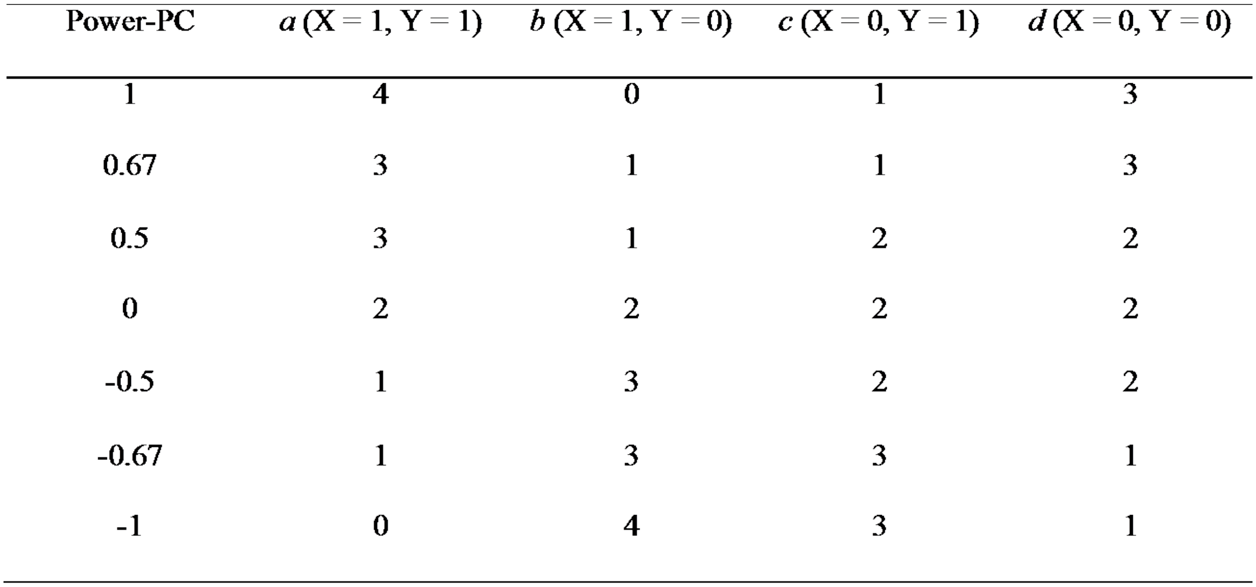
When the steps above did not sufficiently reduce the number of decreases on [A to D] transitions, the cover story was changed to the one reported, where different morning activities have a causal effect on mood. To go along with this, the response scale was changed so that the negative/positive dimension of the cause did not map on to whether the presence of the cause was excitatory or inhibitory, but whether one state over another had a stronger excitatory effect.

Changing the framing of the cause so that its states are not absent vs. present but one activity vs. another changed the interpretation of [A to D] transitions. With the initial interpretation, the large number of decreases on this transition suggested that subjects may have been tracking the effect – it goes from present to absent, and so subjects give a negative rating on the scale. With the new interpretation, subjects observe that the new activity leads to a good mood, and then the regular activity leads to a bad mood. In this case, the subjects are forced to consider how the [D] state is evidence counting against the original activity, which is inconsistent with giving a negative rating.

This format led to the lowest rate of subjects showing this misinterpretation, but it could not be eliminated entirely. This suggests there is some implicit mapping of magnitude to the original vs. new activity, perhaps due to one being mentioned before the other.

APPENDIX C

Table C1. *Frequencies of states in data from Figure 1. Contingencies and corresponding distribution of states used in different scenarios in the experiment.*



Power-PC	*a* (X = 1, Y = 1)	*b* (X = 1, Y = 0)	*c* (X = 0, Y = 1)	*d* (X = 0, Y = 0)
1	4	0	1	3
0.67	3	1	1	3
0.5	3	1	2	2
0	2	2	2	2
-0.5	1	3	2	2
-0.67	1	3	3	1
-1	0	4	3	1

BIBLIOGRAPHY

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative Forgetful Scholars: How People Learn Causal Structure Through Sequences of Interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

Buehner, M. J., & May, J. (2009). Causal Induction from Continuous Event Streams: Evidence for Delay-Induced Attribution Shifts. *The Journal of Problem Solving*, *2*(2), 42–80. doi:10.7771/1932-6246.1057

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. doi:10.1037//0033-295X.104.2.367

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121. doi:10.1016/S0022-2496(02)00016-0

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*(2), 289–344. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10789198

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–84. doi:10.1016/j.cogpsych.2005.05.004

Hagmayer, Y., & Meder, B. (2013). Repeated causal decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 33–50. doi:10.1037/a0028643

Hagmayer, Y., Meder, B., Osman, M., Mangold, S., & Lagnado, D. (2010). Spontaneous Causal Learning While Controlling A Dynamic System. *The Open Psychology Journal*, *3*, 145–162. Retrieved from http://www.benthamscience.com/open/topsyj/articles/V003/SI0088TOPSYJ/145TOPSYJ.pdf

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General*, *138*(1), 22–38. doi:10.1037/a0014585

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: model comparison and rational analysis. *Cognitive Science*, *31*(5), 765–814. doi:10.1080/03640210701530755

Jenkins, H. M., & Ward, W. C. (1965). Judgment of Contingency Between Responses and Outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17.

Johnson, S. G. B., & Keil, F. C. (2014). Causal Inference and the Hierarchical Structure of Experience. *Journal of Experimental Psychology: General*, *143*(6), 2223–2241.

Lagnado, D. A., Fenton, N., & Neil, M. (2012). Legal idioms: a framework for evidential reasoning. *Argument & Computation*, *4*(1), 46–63. doi:10.1080/19462166.2012.682656

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–73. doi:10.1111/cogs.12054

Levin, B., & Hovav, M. R. (1994). A preliminary analysis of causative verbs in English. *Lingua*, *92*(I 994), 35–77. doi:10.1016/0024-3841(94)90337-9

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, *8*(September), 1–12. doi:10.3389/fnhum.2014.00700

Luhmann, C. C., & Ahn, W.-K. (2011). Expectations and interpretations during causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 568–87. doi:10.1037/a0022970

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*(2), 334–52. doi:10.1037/a0014929

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and Intervening: Rational and Heuristic Models of Causal Decision Making. *The Open Psychology Journal*, *3*, 119–135.

Pacer, M. D., & Griffiths, T. L. (2011). A rational model of causal induction with continuous causes. *Advances in Neural Information Processing Systems*, *24*.

Pacer, M. D., & Griffiths, T. L. (2012). Elements of a rational framework for continuous-time causal induction. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In *Classical Conditioning II: Current Theory and Research* (pp. 64–99).

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107–35. doi:10.1111/cogs.12024

Rottman, B. M., & Keil, F. C. (2011). What matters in scientific explanations: effects of elaboration and content. *Cognition*, *121*(3), 324–37. doi:10.1016/j.cognition.2011.08.009

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

Seymour, B., Doherty, J. P. O., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., … Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(June), 664–667. doi:10.1038/nature02636.1.

Shanks, D. R., & Dickinson, A. 1987. Associative accounts of causality judgment. In G. H. Bower (Ed.), *Psychology of learning and motivation: Advances in research and theory* (pp.229-261). San Diego, CA: Academic Press.

Soo. K. W., & Rottman, B. M. (2014). Learning Causal Direction From Transitions With Continuous And Noisy Variables. In P. Bello, M. Guarin, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1485-1490).

Sutton, R. S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, *3*, 9–44.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue Competition in Causality Judgments: The Role of Nonpresentation of Compound Stimulus Elements. *Learning and Motivation*, *25*, 127–151. doi:10.1006/lmot.1994.1008

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian Conditioning: Application of a Theory. In *Inhibition and Learning* (pp. 301–336).

Walsh, C. R., & Sloman, S. A. (2011). The Meaning of Cause and Prevent: The Role of Causal Mechanism. *Mind & Language*, *26*(1), 21–52. doi:10.1111/j.1468-0017.2010.01409.x

White, P. A. (1997). Naive ecology: causal judgments about a simple ecosystem. *British Journal of Psychology*, *88*, 219–233. doi:10.1111/j.2044-8295.1997.tb02631.x

White, P. A. (2014). The Dissipation Effect : A Naive Model of Causal Interactions in Complex Physical Systems. *The American Journal of Psychology*, *112*(3), 331–364.

White, P. A. (2015). Causal judgements about temporal sequences of events in single individuals. *The Quarterly Journal of Experimental Psychology*, *68*(11), 2149–2174. doi:10.1080/17470218.2015.1009475

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111. doi:10.1037/0096-3445.136.1.82

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, *47*(3), 276–332. doi:10.1016/S0010-0285(03)00036-7