Causal Learning with Two Causes over Weeks

Ciara L. Willett (clw137@pitt.edu)
Benjamin M. Rottman (rottman@pitt.edu)
Department of Psychology, University of Pittsburgh,
3939 O'Hara Street, Pittsburgh, PA, 15260 USA

Abstract

When making causal inferences, prior research shows that people are capable of controlling for alternative causes. These studies, however, utilize artificial inter-trial intervals on the order of seconds; in real-life situations people often experience data over days and weeks (e.g., learning the effectiveness of two new medications over multiple weeks). In the current study, participants learned about two possible causes from data presented in a traditional trial-by-trial paradigm (rapid series of trials) versus a more naturalistic paradigm (one trial per day for multiple weeks via smartphone). Our results suggest that while people are capable of detecting simple cause-effect relations that do not require controlling for another cause when learning over weeks, they have difficulty learning cause-effect relations that require controlling for alternative causes.

Keywords: causal learning, multiple causes, trial-by-trial learning, external validity, smartphone

Introduction

When assessing whether or not a potential cause influences an effect, one should control for alternative potential causes. For example, imagine you start trying a new acne medication. When deciding whether or not it works, you should control for other variables that may have changed over time (e.g., sun exposure, diet, stress) that might also affect the probability of acne breakouts. However, doing so could be difficult in that it requires remembering all the variables and deciding on a strategy to control for the other variables.

Accounting for alternative causes is critical for two reasons (see Rottman, 2017; Waldmann & Hagmayer, 2001; for tutorials). If the causes are uncorrelated, controlling for alternative causes reduces noise in the target cause-effect relation and thereby increases the power to detect the relation. If they are correlated, one must account for the other confounded cause when assessing the relation between a single cause and effect; if not, one can draw entirely inaccurate conclusions about the relation between the target cause and effect. Our focus is on this second question.

There are multiple ways to calculate causal strength for a target cause. First, one could calculate the 'unconditional' strength of the target cause (T) on the effect (E) *not* controlling for the alternative cause (A) with Equation 1. In contrast, one could calculate the 'conditional' strength of T on E controlling for 'A' using Equations 2 or 3. This is equivalent to the relation between a single cause and effect in the subset of cases in which the alternative cause is present or the subset of cases in which the alternative cause is absent (e.g., Cheng & Novick, 1992).

$$P(e{=}1|\, t{=}1) - P(e{=}1|\, t{=}0) \quad \text{(Equation 1)}$$
$$P(e{=}1|\, t{=}1, a{=}1) - P(e{=}1|\, t{=}0, a{=}1) \quad \text{(Equation 2)}$$
$$P(e{=}1|\, t{=}1, a{=}0) - P(e{=}1|\, t{=}0, a{=}0) \quad \text{(Equation 3)}$$

Multiple regression accomplishes a similar goal but can handle more causes and also metric variables. In fact, some reinforcement learning algorithms such as Rescorla-Wagner (1972) also control for other cues when calculating associative weights (Danks, 2003). Our study was not designed to test for differences in these theories; all standard approaches will make similar ordinal predictions.

Prior studies have found that people do control for alternative causes (Spellman, 1996; Spellman, Price, & Logan, 2001; Goodie, Williams, & Crooks, 2003). At the same time, people also exhibit a non-rational tendency; when one cause is considerably stronger than the other, participants tend to 'discount' the strength of the weaker cause (Goedert et al., 2005; Laux, Goedert, & Markman, 2010). Another way to explain this is that the stronger cause has a contrastive effect on the assessment of the weaker cause. In sum, there appear to be both rational and non-rational tendencies in assessing the strength of multiple causes.

In the standard trial-by-trial paradigm, participants observe a series of trials, lasting a few seconds each. On each trial, the participant learns if two causes and an effect are present or absent. At the end of the trials, they judge the relationship between each cause and the effect.

However, the standard paradigm is artificial in that the trials are presented very rapidly. We contend that many if not most real-world inferences are made from experiences spanning days, weeks, or even months. When learning is spaced out over time, there are many more distractions and other ongoing cognitive processes. Furthermore, learning over weeks also requires long-term memory instead of short-term memory. Increased demand on verbal working memory has been shown to impede the ability to control for alternative causes (Goedert, Harsch, & Spellman, 2005). In the present study, we studied whether people can control for confounded causes when learning causal relations over longer timeframes.

Prior research suggests that individuals are capable of learning single cause-effect relations after observing one trial per day for multiple weeks (Willett & Rottman, 2019).

However, that task was very simple in that it only involved a single cause and did not require controlling for a second confounded cause. In the current study, we investigated whether people can control for alternative causes by comparing causal judgments for a dataset learned using a traditional trial-by-trial paradigm (24 back-to-back trials) and judgments for the same dataset learned one trial per day for 24 days. This question is practically important as it provides guidance about the accuracy of causal learning in more real-world situations. Furthermore, given that there have been hundreds of studies in causal learning that use the rapid trial-by-trial paradigm and thousands of studies in related fields that employ similar paradigms, this research also has important implications for the external validity of rapid trial-by-trial learning.

Methods

Participants

205 participants (mainly undergraduate students) were recruited for the study; the main requirements were owning a smartphone and intending to complete the 24-day study. Participants were paid $30 if they successfully completed the entire study. The final analyses included data from 191 participants, after excluding 4 people who missed more than three days of the study, 1 person due to potential confusion with the task because they were not fluent in English, 8 people due to a programming or data collection errors, and 1 person who did not show up to their return appointment.

Stimuli and Design

Participants were randomly assigned to learn about one of four datasets that manipulated the 'unconditional' (not controlling for the other cause) and 'conditional' (controlling for the other cause) statistical relations between the target cause, alternative cause, and the effect. The study had a 2 × 2 design in which the target and alternative causes either had a positive (+) or negative (-) conditional influence on the effect, controlling for the other. Table 1 shows the unconditional and conditional statistical relations of the target on the effect using Equations 1 (unconditional) and Equations 2 and 3 (conditional), and analogous statistical relations for the influence of the alternative on the effect. Table 2 shows the number of trials of each type.

Table 1: Summary statistics for the four datasets for the target (T), alternative (A), and effect (E).

Cause-Effect	Dataset			
Statistical Relation	T+A+	T+A-	T-A+	T-A-
T Unconditional	.00	.00	.00	.00
T Conditional	+.33	+.33	-.33	-.33
A Unconditional	+.50	-.50	+.50	-.50
A Conditional	+.67	-.67	+.67	-.67

The unconditional influence of the target on the effect was 0 in all datasets, which means that if participants fail to control for the alternative when assessing the relation of the target on the effect, they would always infer that there is no influence. However, if they do control for the alternative cause, then they would infer a positive relation for two of the datasets and a negative relation for the other two.

For the alternative, notice that the conditional and unconditional assessment of the alternative cause on the effect are qualitatively the same for all four datasets. This means that if participants can learn about the alternative at all, regardless of whether or not they control for the target, they should infer a positive relation for two of the datasets and a negative relation for the other two. Because they do not need to control for the target when assessing the alternative, and also because the alternative is simply stronger than the target, it should be easier to learn about the alternative cause.

Table 2: Number of trials for each combination of target (T), alternative (A), and effect (E) in the four datasets.

			Dataset			
T	A	E	T+A+	T+A-	T-A+	T-A-
1	1	1	3	3	6	0
1	1	0	0	6	3	3
1	0	1	3	3	0	6
1	0	0	6	0	3	3
0	1	1	6	0	3	3
0	1	0	3	3	0	6
0	0	1	0	6	3	3
0	0	0	3	3	6	0

Procedure

Participants completed the entire study using their smartphones. They first completed an 8-trial practice task. Then, each participant completed a short-term (back-to-back trials) and a long-term (one trial per day) version of the same dataset. Participants were randomly assigned to complete either the short or the long task first. Thus, the design was a 2 task length (short vs. long, within subjects) × 2 target influence on the effect (positive vs. negative, within subjects) × 2 alternative influence on effect (positive vs. negative, within subjects). Preliminary analyses revealed possible order effects, so the results presented here only include the analyses from the first task that participants completed, turning the study into a 2 × 2 between subjects design.

All three tasks involved learning about two medicines. The names, shapes, and colors of the medicines were all different. For the practice task, the effect was arthritis pain. For the short and long timeframe tasks, the effects were dizziness and insomnia, randomized. The instructions stated that the medicines could improve, worsen, or have no influence on the effect and the goal was to infer the influence of each medicine on the effect.

During the long timeframe task, participants received text message reminders at 10am, 3pm, and 8pm if they had not yet completed the task that day.

Within a Trial First, at the beginning of each trial, participants were told whether each cause was present or absent and they pressed radio buttons to verify this information. Second, participants predicted whether the effect would be present or absent. Third, participants were told whether the effect was present or absent, and pressed a radio button to verify this information (Figure 1). Fourth, participants were told to imagine the scene for four seconds until the submit button appeared. In the short timeframe condition, participants proceeded to the next trial. In the long timeframe condition, participants were logged out and unable to observe the next trial until the following day.

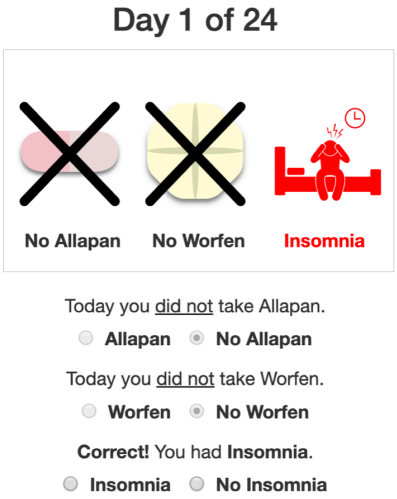

Figure 1. Screenshot from the end of a trial.

Measures of Strength We evaluated participants' beliefs about the strength of the causes on the effect in three ways. First, for each cause, participants made an explicit judgment of "causal strength", calculated from participants' responses to two questions. Participants were first asked whether each cause improved, worsened, or had no influence on the effect. If they said "improved" or "worsened", they answered how strongly the cause improved/worsened the effect on a scale of 1 to 10. These two questions were combined together and mapped onto a -1 to +1 scale such that +1 was that the medicine strongly caused (i.e., worsened) the effect (dizziness or insomnia), and -1 was that the medicine strongly prevented (i.e., improved) the effect. Participants made causal strength ratings after Trial 24 and also before Trials 9 and 17, but only the final rating (Trial 24) is reported here.

Second, after making the causal strength judgment, participants made a "frequency strength" judgment. This involved answering a series of questions in which they were asked how many times they observed each of eight possible trial types. For example, they would recall how many of the past 24 trials in which Medication X was absent, Medication Y was present, and the effect was present. For each participant, we took the average of Equations 2 and 3 to calculate the influence of the target on the effect controlling for the alternative. We used the analogous procedure to calculate frequency strength for the alternative cause.

Third, we created a measure of "predictive strength" from participants' predictions about the presence or absence of the effect on trials 12 – 24 using the average of Equations 2 and 3. We only used Trials 12-24 so that participants would have enough time to learn the relationship. In 59 out of 191 cases, we could only calculate either Equation 2 or 3 because the participant had not experienced all four combinations of the target and alternative in the last 13 trials. In these cases, we used which ever could be calculated.

Results

Our analytical approach closely follows our preregistered plan available at https://osf.io/3dajq/. We first conducted regressions in R for each timeframe condition and measure of strength to evaluate the effects of the target and alternative causes (Table 3). To examine if there were differences between the short vs. long timeframe, we then conducted regressions including the length of the task as a predictor that could interact with the alternative and target cause predictors (Table 4). The regressions code the predictors as +.5 for the positive conditional influence datasets and -.5 for negative. We computed Bayes factors using the BayesFactor package (Morey & Rouder, 2018).

Judgments for the Target Cause (T)

Figure 2 presents descriptive statistics for assessments of the target cause. If participants learned the influence of the target cause on the effect, while controlling for the alternative cause, they would make more positive judgments when the target was positive than when the target was negative. There is some evidence of controlling, especially in the short timeframe condition; this can be seen most easily by comparing the Positive vs. Negative target bars within the Positive Alternative or Negative Alternative conditions.

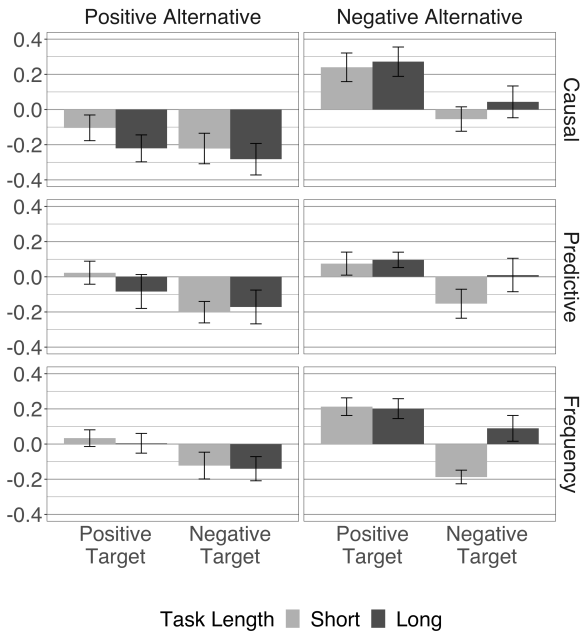

Figure 2: Judgments for Target Cause (SE error bars).

Analytical Approach and Interpretation of Coefficients We conducted regression analyses to test for a main effect of the target (negative vs. positive conditions) and a main effect of the alternative (negative vs. positive conditions) on the ratings of the target cause.

A main effect of target (Tables 3 & 4, Row 1) suggests that participants learned the influence of the target on the effect, while controlling for the influence of the alternative cause. If they did not control for the alternative, there would be no main effect of the target. For this reason, this predictor is called "T (controlling)" in the tables.

A main effect of the alternative cause (Tables 3 & 4, Row 2) tests for the effect of the alternative cause on ratings of the target cause. If there is a significant main effect and the coefficient is negative, this is evidence of non-rational discounting. Thus, this row is labeled "A (discounting)".

Analysis of Short and Long Timeframes Separately The first analysis (Table 3) examined the influence of the actual strength of the target and alternative causes on participants' ratings of the target, separately for the short and long timeframes.

In the short timeframe condition, there was a main effect of the target for all three dependent variables, and it was especially strong for the predictive and frequency strength measures; the BFs align well with the *p* values. This key finding replicates past studies that have found that people are able to control for alternative causes in traditional rapid trial-by-trial designs.

However, in the long timeframe condition, there was not a reliable effect of the target for causal or predictive strength, suggesting that people had difficulty controlling for the alternative when assessing the target cause. There was some evidence of controlling for the alternative in frequency strength (p = .046). However, none of the BFs in the long timeframe condition provide evidence for the null or alternative hypothesis. In sum, though the BFs are not definitive, we do not have any affirmative evidence that participants were able to learn about the target and control for the alternative in the long timeframe condition.

Table 3. 12 Regressions of the 3 measures of strength for the target and alternative, for the short and long timeframes separately.

	Short Timeframe Condition (N=96)									Long Timeframe Condition (N=95)								
	Causal			Predictive			Frequency			Causal			Predictive			Frequency		
Predictor	*b*	*p*	BF	*b*	*p*	BF	*b*	*p*	BF	*b*	*p*	BF	*b*	*p*	BF	*b*	*p*	BF
Target Cause Ratings																		
T (controlling)	.21	**	*	.23	**	**	.28	***	****	.15			.09			.13	*	
A (discounting)	-.26	**	**	-.05		†	-.06			-.41	***	****	-.18	*		-.21	**	***
Alternative Cause Ratings																		
T (discounting)	-.13			.06		†	.09			-.04		†	-.06		†	-.03		†
A (simple learning)	.84	***	****	.96	***	****	.68	***	****	.82	***	****	.75	***	****	.60	***	****

Note: T=Target, A=Alternative.
p-values: *<.05, **<.01, ***<.001.
BFs in favor of alternate hypothesis: *>3, **>10, ***>30, ****>100.
BFs in favor of null hypothesis: †>3, ††>10, †††>30, ††††>100.

Table 4. 6 Regressions of the 3 measures of the target and alternative, with task length as an interaction.

	Causal Strength			Predictive Strength			Frequency Strength		
Predictor	*b*	*p*	BF	*b*	*p*	BF	*b*	*p*	BF
Target Cause Ratings									
T (controlling)	.18	**	**	.16	**	**	.20	*	****
A (discounting)	-.33	***	****	-.12	*		-.14	***	**
L	.01		†	-.03			-.05		
T (controlling) x L	.06		†	.14			.15		
A (discounting) x L	.15			.13			.15		
Alternative Cause Ratings									
T (discounting)	-.08			.00		†	.03		†
A (simple learning)	.83	***	****	.86	***	****	.64	***	****
L	.11			.11	*		.05		†
T (discounting) × L	-.09		†	.13		†	.12		†
A (simple learning) × L	.02		†	.21	*		.08		†

Note: T=Target, A=Alternative, L=Length.
p-values: *<.05, **<.01, ***<.001.
BFs in favor of alternate hypothesis: *>3, **>10, ***>30, ****>100.
BFs in favor of null hypothesis: †>3, ††>10, †††>30, ††††>100.

Row 2 in Table 3 tests the main effect of the alternative cause (i.e., discounting). In the short timeframe, there was an influence of discounting for the causal ratings but not for predictive or frequency measures. In the long timeframe, there were considerable discounting effects for the causal and frequency measures though only a weak effect for the predictive measure.

Testing for Differences between the Short and Long Timeframes In the previous analyses (Table 3) we tested the short and long timeframes separately. We subsequently added the length of the task (L) into the regressions to test if task length moderates the influence of the target or the alternative causes (Table 4). Overall, these analyses revealed effects of the target (controlling) for all three measures and effects of the alternative (discounting) for two of the three measures of strength.

Because these main effects collapse across both timeframes, we were primarily interested in the two-way interactions. Out of the two-way interactions (target × task length, alternative task length), they are mostly nonsignificant and inconclusive in terms of BFs. For causal strength, there was some weak evidence in support of the null hypothesis for the target by task length interaction (BF = 3.47), suggesting that there was no difference in participants' ability to assess the target and control for the alternative between the short and long timeframes.

In summary, there is evidence that participants did control for the alternative when assessing the target cause for the short timeframe, but little (if any) evidence that they controlled for the alternative in the long timeframe. There is evidence of a discounting effect for the causal measure in the short timeframe and for the causal and frequency measures in the long timeframe. Even though there were differences in which predictors were significant for the short and long timeframe, these were not reliably different (2-way interactions in Table 4).

Judgments for the Alternative Cause (A)

Figure 3 presents the descriptive statistics for assessments of the alternative cause. If participants learned the simple relation between the alternative cause and the effect, there would be more positive judgments for the positive alternative dataset than for the negative alternative dataset. This pattern is prominent in Figure 3. If participants discounted the ratings of the alternative cause due to the target cause, then the alternative should be rated relatively lower when the target is positive than negative, which does not appear obviously in Figure 3.

Analytical Approach and Interpretation of Coefficients
A significant positive coefficient for the actual strength of the alternative cause on the assessments of the alternative cause (Table 3 Row 4, Table 4 Row 7) would be evidence for successful learning about the alternative. Note that a positive coefficient for judgments of the alternative cause provides evidence that participants learned about the alternative but does not provide evidence that participants controlled for the target. This is unlike judgments of the target cause, in which a positive coefficient for the target provides evidence that participants controlled for the alternative. The reason is that the conditional and unconditional influence of the alternative cause on the effect are highly similar (e.g., in the T+A+ condition in Table 1, A has an unconditional influence of +.50 and a conditional influence of +.67). For this reason, the main effect of the alternative cause is called "A (simple learning)" in the tables.

A significant negative coefficient of the target cause on ratings for the alternative cause (Table 3 Row 3, Table 4 Row 6) would be evidence for discounting (i.e., discounting the strength of the alternative cause in the presence of the target cause).

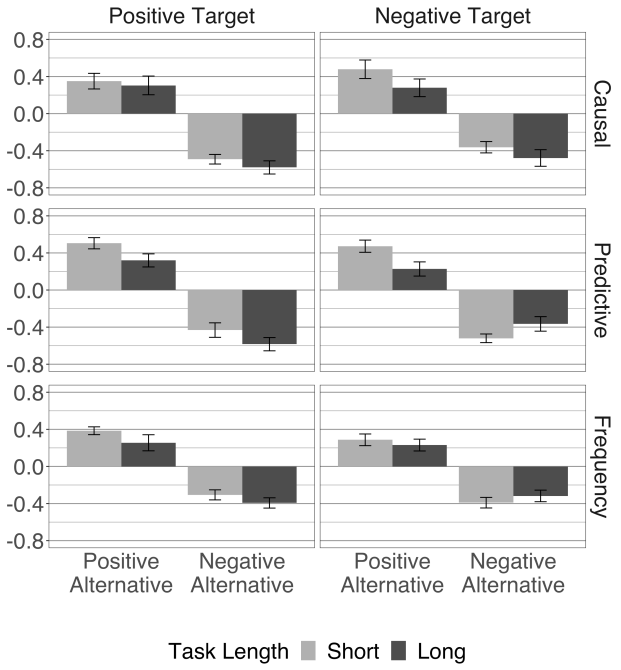

Figure 3: Judgments for Alternative Cause (SE error bars).

Analysis of Short and Long Timeframes Separately
Looking at Table 3, the most notable finding is that across both the short and long timeframes, and across all three measures, there was very strong evidence of a main effect of the alternative cause ("simple learning"). This provides clear evidence that participants can learn about the alternative cause even in the long timeframe.

There were no significant effects of discounting for any of the three measures in either the short or long timeframe. Most of the evidence was in favor of the null hypothesis; no discounting effect. It makes sense that there was no effect of discounting when assessing the alternative cause because the alternative is considerably stronger (both conditionally and unconditionally) than the target. In contrast, when assessing the target, it makes sense that there was more evidence for discounting because the alternative is stronger than the target.

Testing for Differences between the Short and Long Timeframes The bottom half of Table 4 tests for differences between the short and long timeframe when assessing the alternative cause. There is no overall effect of discounting and there are reliable effects of simple learning, which make sense in that these were found for the short and long timeframes separately.

The last two rows of Table 4 tests for interactions between these predictors and timeframe. With regards to discounting (T × L), the length of the task did not moderate the amount of discounting and the evidence was in favor of the null hypothesis. With regards to veridical simple learning of the alternative (A × L), there is some evidence in favor of the null hypothesis that length of task does not moderate learning for the causal strength and frequency strength measure. For the predictive strength measure, there is some evidence that learning was a bit better in the short than the long condition (b = .96 vs. .75 in Table 3).

Discussion

This is the first study testing whether people can learn about two causes over many weeks, and further, whether people can control for alternative causes when learning over weeks. There are three key findings.

First, we found strong evidence that people could learn about the 'alternative' cause in both the short and long timeframe. The alternative cause is fairly easy to learn about because it has roughly the same influence on the effect regardless if one controls for the target or not. The fact that people are able to learn about the alternative in the long timeframe aligns with prior work that found that people are able to learn causal relations between a single cause and effect about as well in the long as in the short timeframe (Willett & Rottman, 2019). Still, the current results build on that finding in that they show that people can also learn about simple causal relations when there are two causes in the long timeframe, not just one.

Second, we found that people could learn about the target cause and control for the alternative in the short timeframe, but we have little (if any) affirmative evidence that they could do so in the long timeframe. This finding is nuanced; we do not have statistical evidence for an interaction with timeframe, so technically it cannot be said that people's assessments of the target cause are better in the short than the long timeframe. Still, there is clear statistical evidence (p values and BFs) that they could accurately learn about the target in the short timeframe, and minimal evidence (just one p-value <.05) that they could in the long timeframe. This is the first piece of evidence that people may have difficulty learning causal relations over weeks and raises the possibility that people may have considerably more difficulty in more challenging situations (e.g., more causes, a delay between the cause and the effect, etc.).

Third, the findings regarding discounting are also important. In the long timeframe condition, the assessments of the target cause revealed strong discounting effects for the causal and frequency strength measures. Remember, in this condition there was no evidence that participants effectively learned about the target cause and controlled for the alternative cause. In fact, the regression weights for the non-rational discounting phenomenon are 1.5 to 3 times stronger than for the rational phenomenon of controlling for the alternative cause (Rows 1 vs. 2 in the right-hand side of Table 3). This raises the concerning possibility that people fail to control for alternative causes in the long timeframe and instead exhibit non-rational discounting tendencies, even though they seem able to learn about simple cause-effect relations that do not require controlling for alternative causes.

Still, the fact that participants' predictions (i.e., predictive strength) exhibited little to no discounting could mean that discounting may be more present in assessments that require more explicit processing. The causal and frequency strength measures involve more sophisticated verbal prompts whereas the prediction task is simpler; furthermore, associative and reinforcement learning models assume that people spontaneously make predictions.

In sum, this research presents important insight as to how people learn causal relations over long periods of time and suggests that people are likely to face challenges when assessing causal relations in real-world situations. Future research should test whether other causal and statistical learning abilities, found in the lab, translate into more real-world situations.

Acknowledgments

This work was supported by NSF 1651330. We thank all of the research assistants who helped with data collection, including Barbara Como, Michael Datz, Isabella Demo, Julia Gillow, Watole Hamda, Beatrice Langer, Marissa LaSalle, Elizabeth Lawley, Brooke O'Hara, and Joanna Ye.

References

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109-121.

Goedert, K. M., Harsch, J., & Spellman, B. A. (2005). Discounting and conditionalization: Dissociable cognitive processes in human causal inference. *Psychological Science, 16*(8), 590-595.

Goodie, A. S., Williams, C. C., & Crooks, C. L. (2003). Controlling for causally relevant third variables. *The Journal of General Psychology, 130*(4), 415-430.

Laux, J. P., Goedert, K. M., & Markman, A. B. (2010). Causal discounting in the presence of a stronger cue is due to bias. *Psychonomic Bulletin & Review*, *17*(2), 213-218.

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of bayes factors for common designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of

reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory, 2*, 64-99.

Rottman, B. M. (2017). The acquisition and use of causal structure knowledge. *The Oxford Handbook of Causal Reasoning,* 85.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7*(6), 337-342.

Spellman, B. A., Price, C. M., & Logan, J. M. (2001). How two causes are different from one: The use of (un) conditional information in Simpson's paradox. *Memory & Cognition, 29*(2), 193-208.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*(1), 27-58.

Willett, C. L., & Rottman, B. M. (2019). The accuracy of causal learning over 24 days. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. Montreal, CA: Cognitive Science Society.