Causal Learning With Delays Up To 21 Hours

by

Yiwen Zhang

B.S., Zhejiang University, 2019

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Yiwen Zhang

It was defended on

September 16th 2021

and approved by

Benjamin M. Rottman, Department of Psychology

Marc Coutanche, Department of Psychology

Scott Fraundorf, Department of Psychology

Causal Learning With Delays Up To 21 Hours

Yiwen Zhang, M.S.

University of Pittsburgh, 2021

Considerable delays between causes and effects are commonly found in real life. However, previous studies have only investigated how well people can learn probabilistic relations with delays on the order of seconds. In the current study we tested whether people can learn a cause-effect relation with delays of 0, 3, 9, or 21 hours, and the study lasted 16 days. Surprisingly, we found that participants were able to learn the causal relation about equally as well in all four conditions. These findings demonstrate a remarkable ability to accurately learn causal relations in a realistic timeframe.

Table of Contents

List of Tables

List of Figures

Preface

This work was supported by NSF 1651330. The author thanks all committee members, Benjamin Rottman, Scott Fraundorf and Marc Coutanche who provided valuable advice to this research, and all of the research assistants who helped with data collection, including Alayna Brothers, Barbaro Como, Shannon Cormier, Micheal Datz, Watole Hamda, Marissa LaSalle, Daniel Lehr, Katherine Lindsay, Elizaneth Lawley, Brooke O'Hara, Lindy Rosen and Alexandria Sitkowski.

1.0 Introduction

The ability to accurately learn cause-effect relations allows people to choose causes that are most likely to bring out desirable outcomes. Hundreds of studies have investigated human causal learning, and many more have investigated correlation detection and reinforcement learning more broadly. However, a critical limitation of this research is that it has investigated learning with relatively short delays on the order of seconds, if any. In contrast, in everyday life, there are many causes that have an influence on an effect a few or even many hours later (e.g., activities during the day that impact the quality of one's sleep, or choices or events one day can impact one's mood the next day). Many people make important decisions based on beliefs about cause-effect relations with delays, but it is not known how accurate such judgments are.

For example, self-diagnosis of non-coeliac gluten sensitivity - fatigue, bloating, and other symptoms related to eating gluten - has become increasingly common (Fasano et al., 2015). However, the medical establishment has difficulty diagnosing non-coeliac gluten sensitivity for a variety of reasons, including that people who think they are gluten sensitive experience strong nocebo effects (experiencing symptoms when taking a non-gluten containing placebo) (Gibson et al., 2017). Deciding to cut gluten out of one's diet has considerable lifestyle implications and is a decision often made on one's own without expert medical guidance (Copelton & Valle, 2009). The question, therefore, is how accurately can people infer that eating gluten-containing foods actually causes these delayed symptoms, and there are countless other analogous questions that people face in daily life.

The goal for the current study was to conduct a well-controlled experiment to understand the role of delay on the ability to accurately detect cause-effect relations.

1.1 Delays in Conditioning and Reinforcement Learning

Temporal delays have been studied in animal conditioning and reinforcement learning for decades. It has often been believed that temporal contiguity is crucial to contingency learning (Renner, 1964; Rescorla, 1967; Skinner, 1948) . Thordike's law of effect (1911) proposed that the connection is more firmly established when the cue is "accompanied or closely followed by" the rewards. Skinner (1948) even posited that contingency can be reducible to the temporal contiguity between the conditioned stimulus (CS) and unconditioned stimulus (US).

Many studies have found that longer delays weaken learning. For instance, in classical trace conditioning, the rate of conditioning is inversely relate to the intra-trial interval (Schneiderman, 1966; Schneiderman & Gormezano, 1964; Smith et al., 1969) . In instrumental conditioning, research over decades has found that animals have more difficulty acquiring a response when the reinforcements are delivered with longer delays (see Boakes & Costa, 2014; Renner, 1964 for reviews).

Because traditional trial-by-trial models of learning do not model intra-trial temporal dynamics, they cannot capture the effects of a delay ("trace learning") (e.g., Rescorla & Wagner, 1972). Subsequent reinforcement learning models such as Sutton and Barto's Temporal Difference model (1990) were specifically designed to capture the negative impact of delay on learning by using a temporally-discounted record of previous events to predict future events, called eligibility traces. Other associative learning theories that are more biologically inspired explain the effect of the delay in trace learning by assuming a short window of associabiability. Those models also encode temporal traces for CSs which determines a short optimal intra-trial interval for neuronal plasticity potential; the potential peaks shortly after the CS and then decreases slowly (e.g., Gluck & Thompson, 1987; Grossberg & Schmajuk, 1989).

However, there are still open debates in the theorizing about the role of delay. First, in a few settings it has been found that learning with delays ranging from 1 hour to 24 hours is still possible (Logue, 1979). This phenomena of learning over a long delay, often called 'preparedness of learning' is believed to be an evolutionarily adapted ability for food-related

conditioned stimuli ('taste aversion') and for certain fearful stimuli and phobias (Dunlap & Stephens, 2014), though is typically not believed to exist in other settings. Second, most of the prior research focused on the time interval between the cue and outcome, but not the inter-trial interval. Gallistel & Gibbon's (2000) timing model proposes a phenomenon called "time-scale invariance" - if the length of delay (response-reinforcer interval) is increased proportionally to the inter-reinforcer interval, then there is no impact of delay. Clearly, how delays affect learning is still a key question (Boakes & Costa, 2014; Gllistel et al., 2019).

1.2 Delays in Human Causal Learning

Within the field of human causal learning, there have also been debates about the role of delay. Initially, it was believed that humans have difficulty learning cause-effect relations with longer delays. For instance, very brief delays have a large influence on the causal judgment of perceptual launching (Leslie & Keeble, 1987; Michotte, 1963; Young & Sutherland, 2009) . Delays longer than 4 seconds significantly reduce causal judgements of action-outcome relations in free-operant conditioning (Shanks, Pearson & Dickinson, 1989).

However, subsequent studies showed that not only are human capable of learning delayed cause-effect relationships, but learning is not necessarily weakened by longer delays. One line of research showed that causal learning is mediated by temporal assumptions (Buehner, 2005; Buehner & McGregor, 2006; Hagmayer & Waldmann, 2002) . Buehner and colleagues argued that Shanks et al.'s (1989) results were due to learners having an expectation of an immediate succession of causes and effects. Buehner and McGregor (2006) found that participants gave stronger causal ratings to a long-delay action-outcome association than short-delay association if they had an expectation of long delay. However, in these studies the longer delays were still on the order of seconds.

Other research investigated the roles of the variability of delay and number of intervening events as opposed to delay per se (Boakes & Costa, 2014; Lagnado & Speekenbrink, 2010), how people use distributions of delays for inferring causal structures among multiple variables (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018) and how people judge singluar

causation (whether one event was the cause of another) based on delay and the presence of other causes (Stephan, Mayrhofer & Waldman, 2020). Those studies indicate that people make use of delays to guide their judgments about causation.

1.3 Longer Delays and Current Study

Despite all this empirical and theoretical work on delay, an important open question is how delays impact human learning in real-life situations. Almost all the prior research, with the exception of the preparedness of learning research with animals, has focused on delays on the order of seconds. However, many real-life causal events occur with delays of several minutes, hours, or days. The primary goal of this research is to investigate how well people are able to learn cause-effect relations with delays on the order of hours.

Recently we have begun studying how well people can learn cause-effect relations from data presented one trial per day for a series of days, which we call 'ecological momentary experiments' (in contrast with ecological momentary assessments). The reason is that one way in which standard causal-learning paradigms are artificial is that all the trials are presented in quick succession, whereas in the real world (e.g., learning if a medicine is working, or what factors influence sleep), the experiences are spaced out over much longer periods of time. We have found that people can learn true relations between a single cause and a single effect about as well when spaced out one trial per day as when presented rapidly within a few minutes (Willett & Rottman, 2021). Furthermore, in both short and long timeframe conditions, participants incorrectly inferred correlations that did not exist ('illusory correlation') when observing skewed datasets. Follow-up research focused on people's ability to learn about two causes and one effect in a long timeframe setting (Willett & Rottman, 2020). Another study investigated reward-based learning, not causal learning, in both a single massed session vs. several sessions spaced out over time. In this study, performance was not different between the massed and spaced conditions when tested immediately after learning, but performance was maintained significantly better in the spaced condition when tested 3 weeks later (Wimmer et al., 2018).

Critically, in all three of these studies there was no delay, so even though they are more realistic in one sense that the trials are more spaced out, they are still artificial in that there was very little delay if any between the cause and effect or action and feedback. The findings that participants are able to learn single cause-effect relations quite well, and are able to learn about two causes to some degree, possibly represent an overly optimistic picture of real-world causal learning. The current study added delays between causes and effects in the long timeframe, which simulates more realistic real world causal leanring.

In the current research, we conducted an ecological momentary experiment aimed to assess the effect of temporal delays on causal learning over 16 days. We adopted a trial-by-trial learning paradigm and spaced it out to one trial per day. We manipulated the intervals between the cause and the effect within a trial, ranging from a few seconds to roughly 21 hours, to investigate whether long term causal learning is impeded with delay and the extent to which people can accurately learn cause-effect relations with long delays.

2.0 Methods

2.1 Participants

202 participants completed the study (150 females, $M_{age} = 22.1$, $SD_{age} = 5.6$). 76 participants were recruited within the Pittsburgh community (mainly undergraduate students) and attended an in-person lab session on the first day of study. Due to the COVID-19 pandemic, the rest of participants were recruited through social media (e.g. Facebook) and attended a video session over Zoom on the first day of study. Participants who successfully completed the entire study were paid $40. The final analyses included 200 participants, excluding 1 participant who explicitly reported they wrote down data during the study and 1 participant due to a programming error.

2.2 Design

The study employed a 2×4 between-subject design. There were two types of learning datasets (positive correlation vs. negative correlation) and four temporal delay conditions of roughly 0, 3, 9, or 21 hours between the cause and effect.

2.2.1 Datasets

In the positive dataset, the cause generated the effect (i.e. taking medicine was associated with pain), and in the negative dataset, the cause prevented the effect (i.e. taking medicine was associated with no pain). The positive correlation dataset used the following data: the cause and effect were both present 6 times (A cell), both absent 6 times (D cell), the cause was present and the effect was absent 2 times (B cell), and the cause was absent and the effect present 2 times (C cell). For the negative dataset, the cell frequencies were reversed [A=2, B=6, C=6, D=2]. According to the ΔP rule (Allan, 1980), the contingency between

cause and effect were .5 and -.5 for the two datasets respectively. According to the Power PC rule (Cheng, 1997), the causal power was +.66 and -.66. In order to ensure that the contingency was kept the same for the first 8 days and the latter 8 days, we divided the whole dataset into two identical sets ([A=3, B=1, C=1, D=3] for the positive; [A=1, B=3, C=3, D=1] for the negative) and randomly ordered 8 trials within each half.

The reason for testing two different datasets was simply to distinguish learning from a bias (e.g., a bias that participants on average believed that the medicine would be helpful or harmful). Our goal was not to compare learning for the two conditions.

2.2.2 Temporal Delays

We manipulated the temporal delays within each trial. Participants observed 16 trials and each trial contained a cause task in which participants learned whether the cause was present or absent, and an effect task in which participants learned whether the effect was present or not. In the 0-delay condition, participants did the cause and the effect task back to back each day. In the 3-hour delay condition, participants did the cause task in the morning and the effect task in the afternoon around 3 hours (min = 2, max = 7) later than the morning task. In the 9-hour delay condition, participants did the cause task in the morning and the effect task in the evening around 9 hours (min = 8, max = 15) later. In the 21-hour delay condition, participants did the cause task in the afternoon and the effect task the next morning roughly 21 hours (min = 16, max = 24) later. See Figure 1 for a visualization of these windows of time to participate and Figure 5 for the distributions of average time intervals that participants actually experienced over 16 days.

The study was run automatically through a custom built website using the psychcloud.org framework. This website sent automated text message reminders, and allowed participants to login only at the allocated times. When participants were supposed to do the task, they were sent a text message, and if they did not do the task they received hourly reminders.

If a participant did not do one of the tasks (either the cause or effect task) within the window of time that they were allotted on a given day, they were not allowed to participate for the the rest of the day, and they received the same trial the subsequent day. This means

that sometimes the cause task was repeated from one day to the next, but the effect task for a given trial was never repeated, so there was only one opportunity to learn about the cause-effect relation in a given trial. If a participant missed more than 4 days, they could not continue to participate in the study. In total 6 participants were dropped from the study due to missing more than 4 days.

2.3 Procedures

The entire study was conducted on participants' mobile phones. The study contained one practice task which happened in the lab (or over Zoom) on Day 0, one 16-day learning task and one final judgement task which happened on Day 17.

On the first day (Day 0), participants were introduced to the study and did a practice task to gain familiarity with the procedure. The practice task contained a four-trial learning session and a testing session afterwards. In the learning session, the cause and effect tasks were completed back-to-back.

The long-term task began on Day 1. At the beginning, the participants read a cover story designed to make it plausible that the medicine could improve or worsen the outcome, as follows:

"Please imagine that due to a health condition, you are on a medication called Primadine. In addition to that health condition, you also sometimes have pain from arthritis. You have heard that sometimes Primadine can improve or worsen the pain as a side effect.

Some medications happen to improve arthritis pain as a side effect by decreasing the autoimmune processes that cause inflammation and pain in arthritis.

Other medications happen to worsen arthritis pain as a side effect by increasing the autoimmune processes that cause inflammation and pain in arthritis.

For 16 days you are going to see whether or not you take Primadine and whether or not you experience pain.

You want to figure out whether Primadine improves or worsens or has no influence on your pain."

The entire learning session contained 16 trials, one trial per day. Each day participants conducted two tasks, a cause task and an effect task (see Figure 2).

In the cause task, participants first saw an image of a scene. After they clicked the 'Continue' button, they were shown an icon and text of whether the cause is present or absent that day. Participants then verified whether the cause was present or absent by clicking a button. Only after they responded correctly would a 'Continue' button appear allowing them to continue. Finally, they were asked to "tell a story that links both pictures together."[1]

The effect task followed a similar procedure, except that before seeing whether the effect was present or absent, participants were asked to predict the status of the effect (whether or not they have back pain). For the prediction, they were not reminded whether the cause was present or absent, and the cause was not mentioned at all. After they submitted their prediction, they received text feedback of their prediction and an icon showing whether they had back pain or not, verified the presence or absence of the effect, and also wrote a story linking the effect and contextual image.

On Day 17, the day after the 16-day learning task, they did a 15-minute final judgment task. The task consisted of two parts. First, participants made four judgments of the cause-effect relation. Second, participants were asked to recognize the contextual images they saw each day and recall whether or not the cause and effect were present based on the images. We will not discuss the memory task further in this article.

[1]We wanted to ensure that participants were paying attention and encoding the stimuli, not just clicking through the task, given that this task was embedded in their daily lives and could happen while they were doing other things. We were also concerned that if learning in all conditions was at floor it could be explained merely due to a lack of processing. One potential concern is that this task may have led to increased salience, perhaps leading to an overly optimistic picture of learning with delays. Though possible, these stories were still quite short and likely took 10-20 seconds to write. In comparison, many real-world events that people care to learn about are likely to be much more salient and important in one's life (e.g., pain, sleep) leading to deeper processing than in the current task.

2.4 Measures

We used four measures of participants' beliefs about the strength of the relation between the cause and the effect at the end of the study. We also converted the trial-by-trial predictions of the effect during the learning phase into a measure of learning. All the measures were scaled in a range of [-1,1] for analysis.

2.4.1 Causal Strength

Participants made a standard "causal strength" judgment by answering "Do you think that Primadine worsens, or improves pain?" (on a scale of -10 = strongly worsens, 0 = no influence, to $+10$ = strongly improves). This question was asked both in the middle of the learning session (before Trial 9) and in the testing session (after Trial 16).

2.4.2 Future Prediction Strength

Participants were asked about the probability of having pain given that they did or did not take the medicine with the following question: "Imagine that 'tomorrow' (Day 17) you take/do not take Primadine. On a scale of 0 to 100%, what do you think is the likelihood that you would experience pain?" The future prediction strength was derived by subtracting participants' responses of when they do not take the medicine from when they do take the medicine - similar to the ΔP rule (Allan, 1980).

2.4.3 Future Use Strength

Participants answered "Do you think you should continue to use the Primadine" on a scale of -10 = definitely no, 0 = unsure, to $+10$ = definitely yes.

2.4.4 Frequency Strength

We asked about participants' memories of the frequencies of A, B, C, and D cells (e.g., for the A cell we asked "Of the 16 days in the study, how many days did you see a picture in

which you did take Primadine and did experience pain"). We calculated frequency strength by calculating $p(\text{effect} \mid \text{cause}) - p(\text{effect} \mid \neg\text{cause})$ from participants memories of A, B, C, and D cells. We excluded one participant from data analysis due to this participant's frequency strength being incalculable due to a division by zero problem, which can happen if some pairs of cells are judged as zero.

2.4.5 Trial by Trial Prediction Strength

We computed "trial by trial prediction strength" from participants' predictions about the presence or absence of the effect from Trial 9 to Trial 16 using the following equation: $p(\text{predicted effect} \mid \text{cause}) - p(\text{predicted effect} \mid \neg\text{cause})$. We only used the last 8 trials to have a measure after participants had an opportunity to learn the cause-effect relation, not right at the beginning of learning.

Figure 1: The time windows for participation.

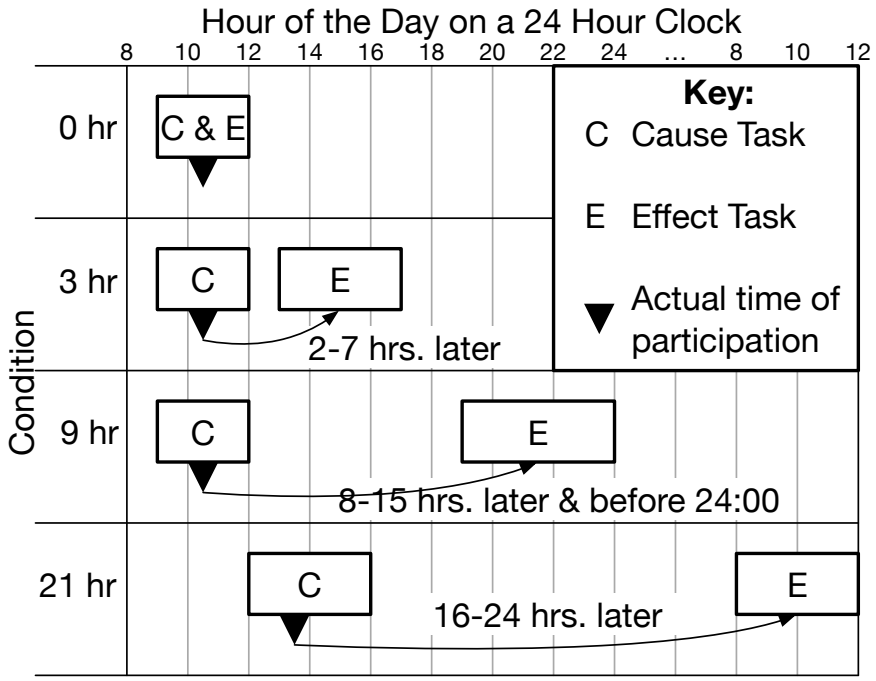


Figure 2: Screenshots of the cause task and the effect task.

Note. In the cause task (left), participants learned and verified whether the cause was present or not based on the icon (whether or not they took the medicine), and then wrote a story to related the contextual image and the cause icon. After 0, 3, 9 or 21 hours delay, in the effect task (right), participants predicted that the effect would be present or absent (pain or no pain), received feedback, and verified the state of the effect. At the end of the effect task, participants also wrote a story to relate the contextual image and the effect.

3.0 Results

The analysis follows our pre-registered plan available at https://osf.io/8tcvq. For ease of interpretation, we inverse coded the strength judgements for the negative datasets so that they are positive, and show them both together after reverse coding in Figure 3. As can easily be seen in the Figure 3, participants' judgments were above zero for all six measures and for all 4 delay conditions, which provides evidence that participants were able to learn the contingency between the cause and effect in every condition. All 24 t-tests against zero were significant (see the t test labels in Figure 3), which provides evidence that participants learned the contingency above chance in every condition. The all but one of the p-values were ¡.001 and the Cohen's D's were in the range of 0.43 to 1.21. We conducted two analyses to test the influence of delay. One compared the four delay conditions using an ANOVA. The other analysis was a linear regression testing the influence of the actual time intervals that a participant experienced (see the distribution of the actual time intervals in Figure 5). Because the two analyses are very similar, we report the ANOVA results in the main paper and the regression analysis in Appendix as well as mention the few discrepancies in the main paper.

The ANOVAs tested for main effects of delay, dataset (positive vs. recoded negative) and an interaction. If learning becomes weaker with longer delays, there would be a main effect of delay.[1]

Table 1 presents the ANOVA results. We first discuss the main effect of delay. In four of the measures (causal strength before Trial 9, and after Trial 16, future use strength, and future prediction strength), there was no significant effect of delay, and the Bayes Factors (BFs) were in the range of .03-.06, meaning that the evidence is roughly 20 to 1 in favor of the null hypothesis of no influence of delay. In the frequency strength measure, the p-value was

[1]There are two highly related ways to conduct this analysis. One way involves testing for an interaction between dataset and delay; if participants have more difficulty learning the cause-effect relations then their judgments for the positive and negative datasets would get closer together over longer delays. Here as preregistered, we took a simpler approach of inverse coding the judgments for the negative datasets so that they are positive and then testing for a main effect of delay. These two approaches are very similar mathematically and reach the same conclusions, only here we are primarily interested in a main effect of delay whereas in the other version we would primarily be interested in the interaction.

still non-significant and the BF was weaker, about 2 to 1 in favor of a null effect. (Different from the ANOVA result, in the regression analysis for the frequency strength measure, the p-value was .035 showing a significant effect of delays, but the BF was only 2.07 in favor of an influence of delay.[2])

The only measure that found a reliable influence of delay was the trial-by-trial prediction strength measure (p=.006, BF=5.13 for the ANOVA and p=.003, BF=13.6 for the regression). As can be seen in Figure 3, participants' predictions were stronger for the 0 and 3 hour conditions than for the 9 and 21 hour conditions, suggesting that learning was somewhat better with the shorter delays. We compared the trial-by-trial prediction strength measure across the 0 vs. 3, 3 vs. 9, and 9 vs. 21-hour delay conditions. Out of these three comparisons, the only significant difference was 3 vs. 9 hours, $F(1,95) = 7.048$, $p = 0.009$, BF = 4.8.

In the above analysis the trial-by-trial prediction strength was computed from Trials 9-16 to capture behavior after participants had already had an opportunity to learn the relationship. Figure 4 plots learning curves using trial-by-trial prediction strength for the trials binned into groups of 4 collapsing across participants. As can be seen in the figure, the biggest differences among four conditions occurred between Trials 9 to 12; by Trials 13-16 the differences between conditions were smaller.

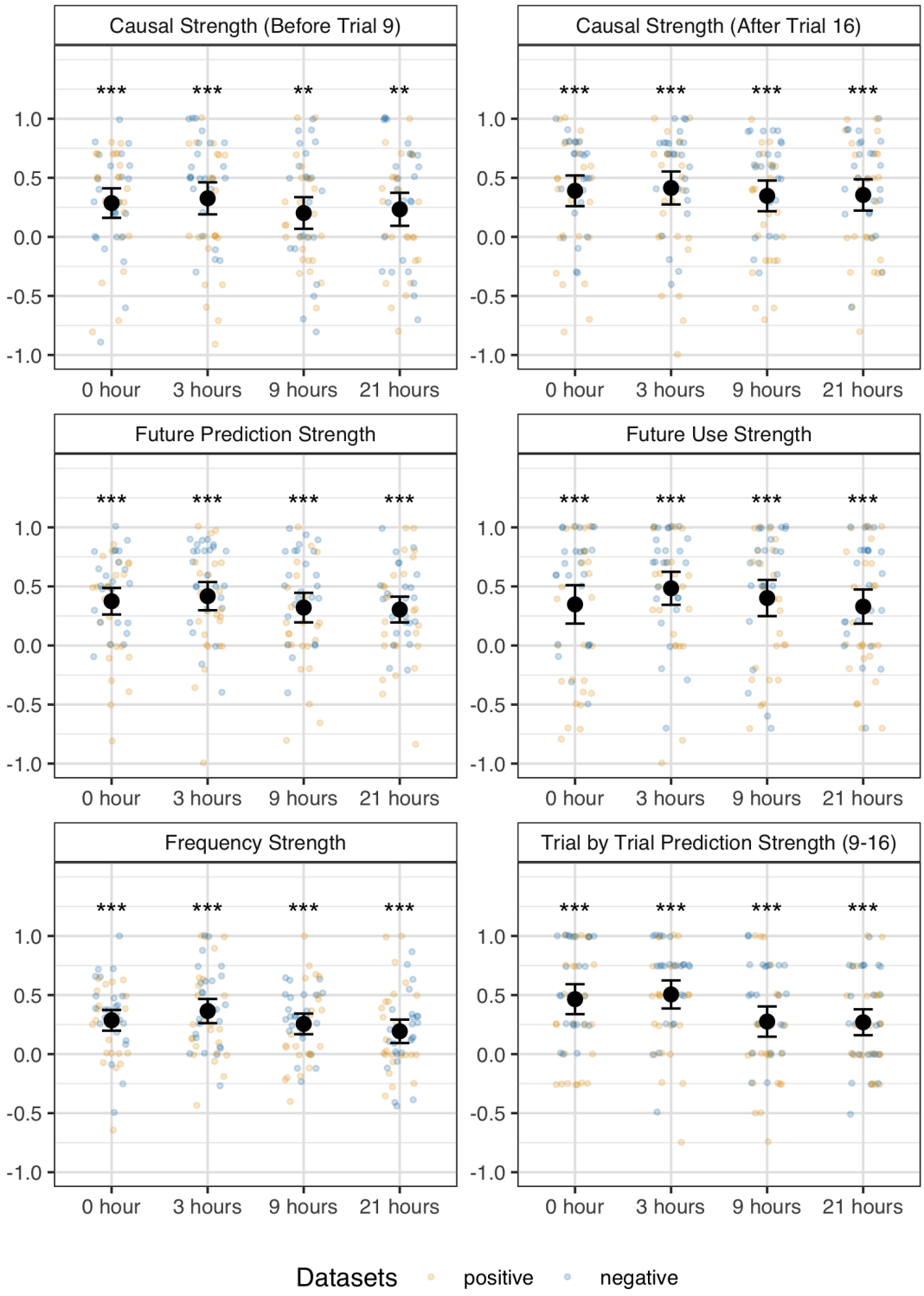
In sum, four of the measures found evidence for a null effect of delay. The frequency strength measure was more inconclusive, and the trial-by-trial prediction strength measure did find some evidence for an influence of delay.

We now move on to the other main effect and interaction. Four out of the six ANOVAs also found very strong main effects of dataset. This reflects the fact that participants tended to give somewhat stronger judgments in the negative condition than the positive condition. Because the negative condition was reverse coded, the main effect means that the recoded judgments in the negative condition are more positive than the judgments in the positive condition. This could have been due to a bias to think that the medicine is effective - that the presence of the medicine would help prevent the back pain. Though the cover

[2]This is an example of how significant p-values especially in the range of .01-.05 can have weak BFs (Wetzels et al., 2011).

story explicitly said that the medicine could improve or worsen the back pain, this bias is understandable, and is not of primary importance to the study. There were no significant interactions between delay and dataset, with most of the BFs roughly in the range of 10 to 1 in favor of the null.

Figure 3: Mean and 95% confidence interval of causal judgments in four delay conditions and t-test results.



Note. *p-values against 0*: *<.05,**<.01, ***<.001.

Figure 4: Trial by trial prediction strength for every 4 trials.

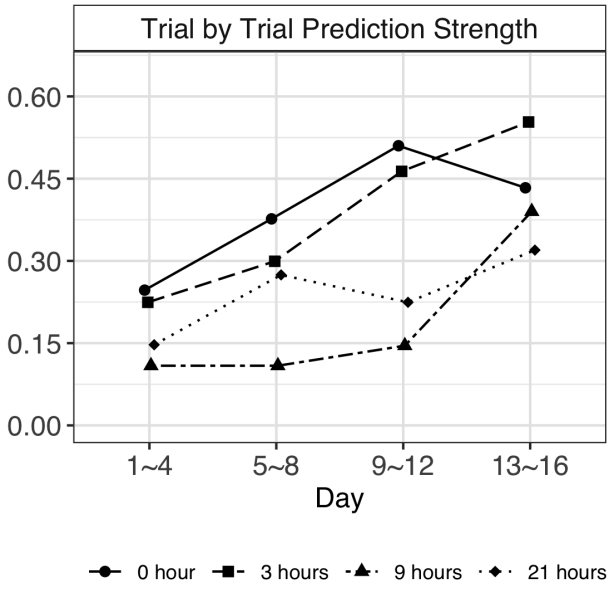


Table 1: ANOVA results for Causal Judgment measures.

	F	p	BF	η_p^2
Causal Strength (Before Trial 9)				
Delay	0.62	0.603	0.05	<0.01
Dataset	3.98	0.048	1.03	0.02
Delay:Dataset	1.13	0.340	0.19	0.02
Causal Strength (After Trial 16)				
Delay	0.18	0.908	0.03	<0.01
Dataset	14.17	<0.001	>100	0.07
Delay:Dataset	0.27	0.846	0.08	<0.01
Future Prediction Strength				
Delay	0.76	0.518	0.06	0.01
Dataset	14.75	<0.001	>100	0.06
Delay:Dataset	0.57	0.635	0.10	<0.01
Future Use Strength				
Delay	0.85	0.467	0.06	0.01
Dataset	12.57	0.001	55.71	0.06
Delay:Dataset	0.36	0.783	0.08	<0.01
Frequency Strength				
Delay	2.24	0.085	0.41	0.03
Dataset	1.65	0.200	0.34	<0.01
Delay:Dataset	0.46	0.709	0.09	<0.01
Trial by Trial Prediction Strength				
Delay	4.28	0.006	5.13	0.06
Dataset	12.12	0.001	45.9	0.06
Delay:Dataset	0.78	0.510	0.13	0.01

4.0 Discussion

Understanding how well people can learn cause-effect relations with considerable delays is crucial to understanding how accurately people can learn cause-effect relations in the real world. This is the first study that used a controlled experiment to test causal learning with delays on the order of hours rather than seconds.

Our approach for studying causal learning in more real-world timeframes is to incrementally move towards more and more realistic learning situations. In our first study on learning over a long timeframe, we simply stretched out the standard trial-by-trial paradigm so that one trial occurred each day (Willett & Rottman, 2021). In the current study, we took the next step of introducing delays between the cause and effect. We have also studied learning over long timeframes with two causes but with no delays (Willett & Rottman, 2020). We are intentionally taking small steps to understand how individual changes affect learning.

Running against the standard assumption that even short delays considerably impair learning, we found that for the most part people were able to learn cause-effect relations with delays up to 21 hours with minimal impact of the delay. There was some evidence of an influence of delay in the trial-by-trial predictions and potentially the frequency strength judgments, but not the other measures.

We were surprised how well participants performed in the delay conditions. First, it was not obvious that people would have strong memories for whether or not the cause happened 9 or 21 hours earlier. If they forgot or got confused about whether the cause occurred earlier in the day, this noise would have presumably impaired learning. Indeed, even in short timeframe learning situations, people are worse when their working memory is taxed (e.g., Kao & Wasserman, 1993; Goedert et al., 2005; Aitken et al., 2001; De Houwer & Beckers, 2003; Liu & Luhmann, 2013; Waldmann & Walker, 2005) and age-related memory decline is associated with worse learning (Mutter et al., 2012). We had thought that delays between 0 to 21 hours would reveal a considerable decrement to learning, and still think that it is possible that if the same paradigm were extended to delays of 2 or 3 days that at some point there may be a decrement, especially if the delay is long enough such that there are other

cause or effect events in-between when an individual cause has an impact on the effect. For example, if a cause influences an effect two days later, but in between there are two causes and one other effect, this might be very hard to learn. Still, finding a lack of an influence of delay up to 21 hours is still important in that it helps to narrow down when we can expect real-world causal judgments to be accurate or not accurate.

Second, even if they did remember whether the cause happened earlier in the day, it is not a given that people would have relatively accurate causal assessments at the end of 16 days. We have found that in the long timeframe, similar to the short timeframe, people tended to infer illusory correlation when observing skewed datasets (Willett & Rottman, 2021), so causal learning is clearly imperfect. We also assume that when the causal learning task gets hard enough (a long enough delay, too many causes, or other complex dynamics) that people will have a very difficult time learning the relations. For example, whereas associative learning and reinforcement learning approaches only require updating a belief about the cause-effect relation each day, 'rule-based' models of causal inference (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Hattori & Oaksford, 2007) require people to have memories of tallies of the all the experiences, and if these tallies are inaccurate the final judgments would also be inaccurate. We found a bit of evidence of worse memories of such tallies for the frequency strength measure in the regression analysis though not in the ANOVA analysis, and more broadly, the causal judgments appeared to be mainly intact across all levels of delay.

4.1 Potential Explanations for Inconsistencies in Findings

We used a number of dependent measures; most showed that learning was not impacted by the delay, though there were some exceptions. Most notably, the length of delay affected the trial-by-trial prediction strength. There are a number of potential factors that could have contributed to influence of delay appearing primarily for the trial-by-trial predictions and not the other measures, though unfortunately none of the available explanations proposed below is entirely satisfactory.

4.1.1 Interim vs. Final Measures of Learning

Unlike the other measures which were collected at the end of learning, the trial-by-trial predictions of the effect were collected during each effect task during learning, so this measure captures interim learning. Could this explain the difference? Figure 4 shows that the largest impact of delay for the trial-by-trial measure was roughly during Trials 9-12. However, there was not an effect of delay in the causal strength measure collected before Trial 9, suggesting that learning was not considerably impaired by the delay half way through learning. It is still possible that the trial-by-trial measure is more sensitive or is picking up differences appearing earlier or later than the causal strength measure half way through learning, but the lack of an influence of delay for the causal strength measure half-way through learning provides some evidence that the difference may not be due to the interim nature of the trial-by-trial predictions.

4.1.2 Direct vs. Indirect Assessments of the Cause-Effect Relation

Whereas some of the measures like causal strength and future use strength are fairly direct measures of the cause-effect relation, the trial-by-trial predictions of the effect is a less direct measure. Perhaps the less direct questions are more susceptible to an influence of delay? On the one hand, the frequency strength judgment, which is also less direct, also showed some (though less convincing) evidence of an influence of delay. The frequency strength measure had people to recall their memories for the 16 events rather than make a summary judgment about the relation between the cause and effect. On the other hand, no influence of delay was seen in the future predictive strength measure, which was based on participants' predictions of the effect given the cause on a hypothetical 17th day. The trial-by-trial prediction strength and future prediction strength measures are very similar to one another, which makes this explanation less convincing.

4.1.3 Effort Required to Recall the Cause to Make Predictions

When making the predictions of the effect, participants were not reminded of the state of the cause that was presented earlier in the day - so accurate predictions would have required participants to spontaneously recall whether the cause was present or absent earlier and then use that information to predict the effect given their belief about the statistical relation between the cause and effect. It is possible that this requires more effort in the longer delay conditions so that participants may have been less likely to make predictions of the effect based on the cause they saw earlier.

However, the notion that people in the long delay conditions did not spontaneously recall the cause is problematic from all learning perspectives. Rule-based theories of learning assume that people keep tallies of the different types of events of the cause and effect - if people do not spontaneously recall the cause and encode it with the effect, then such accounts would not be able to explain the successful learning measured by the other dependent measures. One could argue that participants do not necessary need to spontaneously recall the cause when asked to predict the effect, so long as they do so immediately after when presented with the effect. Though possible, it is not clear why people would only spontaneously think of the cause when presented with the effect, not when asked to make a prediction that would be facilitated by thinking about the cause.

This notion is also problematic from a reinforcement learning perspective. Reinforcement learning models (e.g., Rescorla & Wagner, 1972; Sutton & Barto, 1998) assume that learners spontaneously make predictions of outcomes based on the cues, and that these predictions are required for learning because learning is driven by prediction error. Thus, according to these models difficulties making predictions should also be seen in difficulties in final judgments. However, participants' final judgments did not show an influence of delay.

4.1.4 The Impact of Repeated Predictions

We had participants make trial-by-trial predictions of the effect simply to increase attention to the task. There is not a lot of work on the impact of including predictions in sequential learning tasks. Well et al. (1988) found some hints that requiring trial-by-trial

predictions may have improved discrimination for weak correlations at the end of learning, though the potential effect was small. A benefit of making predictions could be viewed as fitting with the retrieval practice and testing effect literature that finds that repeated practice retrieving to-be-remembered information improves long-term retention (e.g., Roediger & Karpicke, 2008; Rowland, 2014). At the same, standard reinforcement-learning theories assume that people spontaneously make predictions and that learning occurs through prediction error. If people already spontaneously make predictions perhaps explicit predictions may not have much of an impact.

There is some literature that embedding other types of judgments during learning, not just at the end of learning, may impact final judgments. Collin and Shanks (2002) found that embedding causal judgments during the learning led to final judgments that were more aligned with the most recent evidence (a recency effect) whereas when participants only made a single final judgment at the end of the learning without making interim judgments their judgments were more consistent with the overall dataset or even primacy (see also Hogarth and Einhorn, 1992, pg 16; Marsh & Ahn, 2006; Dennis & Ahn, 2001 for related theories and evidence). However, all four delay conditions included the trial-by-trial predictions and one intermediary causal strength judgment, and the primacy vs. recency effect is not relevant to the stimuli we used. In sum, this finding does not explain why the trial-by-trial predictions revealed an effect of delay but other measures did not.

4.2 Open Questions

There are a number of open questions. First, in the current study there was only one cause and one effect, but in many real-life situations there may be multiple causes and multiple effects and various delays between them which could complicate learning. Some research has shown that when there are multiple causes, it is not the delay per se that leads to worse learning but instead it is the number of intervening events or both (Boakes & Costa, 2014; Lagnado & Speekenbrink, 2010; Revusky, 1971). Still, when there are multiple causes that can occur at different times, relations with longer delays will typically have more intervening

events, presumably making them harder to learn. An interesting point to consider is that in the current study, the conditions with longer delays actually had many more intervening events - it is just that these events occurred outside the context of the smartphone app. Even though it may be relatively easy to filter out events deemed to be irrelevant in this study simply due to the context of the phone, in real world situations people would need to more actively choose to focus on certain events and not others, potentially making delay harder. The current studies may suggest that a top-down attention-driven process that filters out intervening events deemed to be irrelevant may be important for accurate learning with delays. This sort of mechanism of what intervening events are 'relevant' (e.g., Revusky, 1971) is outside the bounds of existing theories of causal learning and reinforcement learning.

Second, the current study used a trial-by-trial classical or Pavlovian paradigm in which the participants were presented with the cause and effect or the absence of the cause and effect. In this sort of paradigm it is easy to map the single incidence of the cause to the single incidence of the effect. However, in free operant instrumental paradigms, the participant can choose to implement the cause(s) repeatedly. This might reveal more of an influence of delay because it would be hard to map a single causal event to a single effect event - it is possible for there to be many causal events before an effect event or an effect event without an obvious recent causal event. Shank's et al's (1989) finding of worse learning with delays on the order of seconds utilized a free operant paradigm. At the same time, Gallistel argues that in this situation, it is not the absolute delay between the cause and effect that matters, but rather the ratio between all the events including the inter-trial interval - he argues that when all of aspects of the study are proportionally stretched out in time, learning proceeds at the same pace, called timescale invariance (Gallistel & Gibbon, 2000) . We are planning on testing this hypothesis in a free operant paradigm using long delays of multiple hours like in the current study.

Third, in current study, participants learned only a single cause and a single effect, and both the presence and absence of the events were very salient. However, in many real life situations, nothing alerts the learner to an absence, which raises important challenges for research (Gallistel, Craig, & Shahan, 2019, Hattori & Oaksford, 2007). For example, if someone forgets to take medicine, or does not feel pain, these absences may go unnoticed.

However, modifying the paradigm for real absences will be hard to implement because if a participant fails to do a task when they receive a text message it will turn what should be a presence into an absence.

Fourth, another important future direction has to do with a different sort of delay. Some delays do not happen on a trial basis. For example, antidepressant medications can take multiple weeks before starting to have an influence and in this case the medicine is taken each day, whereas in the current study the medicine was taken on 50% of days. We are also investigating this other sort of delay in a separate line of research.

In sum, it is possible that the reason we did not see much of an influence of delay was that the study was still too simple. At the same time, some argue that stretching learning out over time should not affect learning even in more complex paradigms. We took an approach of making one change at a time to existing paradigms to investigate progressively more complex and realistic learning situations, so that we can hopefully identify the specific factors responsible for challenges in learning.

5.0 Conclusions

This research makes an important empirical contribution to the field of human causal learning specifically, and learning more generally, showing that learning is not necessarily degraded even with considerable delays. Empirically, this raises the possibility that people can accurately learn about the contingencies between events in their daily lives, at least in simple cases with only one cause and effect. Theoretically, this research requires a reexamination of the mental processes that underlie human causal and statistical learning, which primarily assume that learning is degraded with increasing delay. Still, it is important to try to test more complex and realistic learning situations, which may yet reveal impacts of delay.

Appendix A. Regression Analyses

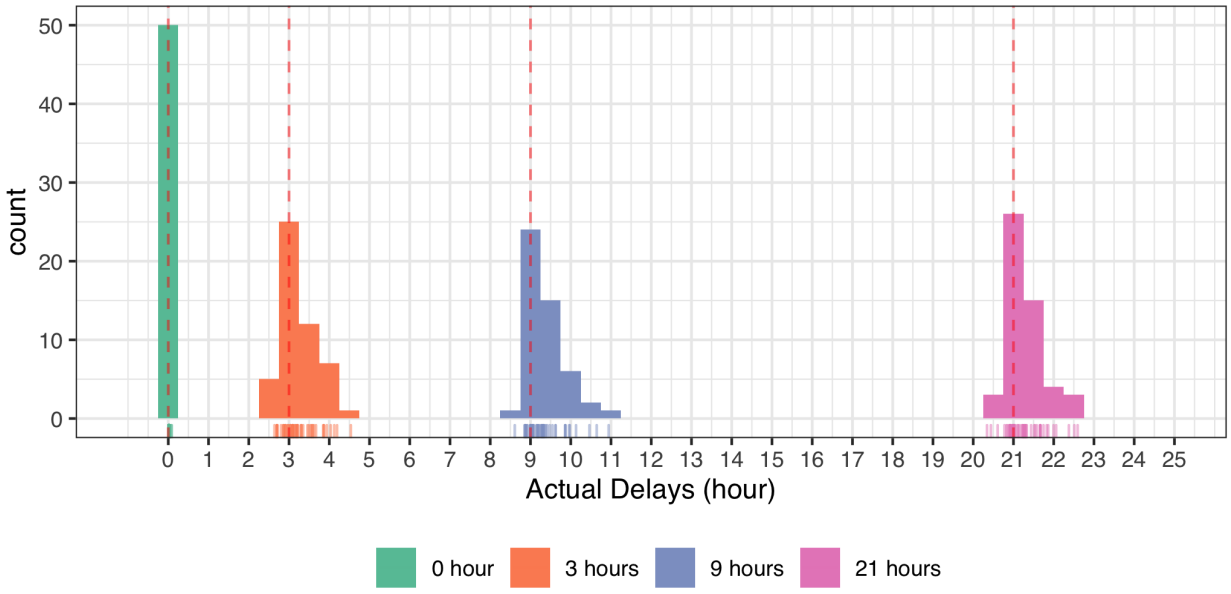
Addition to the ANOVA analysis, we conducted linear regressions to test for an effect of the actual time intervals that a participant experienced. This analysis also differs from the ANOVA analysis in that it tests for a linear influence of delay rather than merely any differences between the four conditions.

In the experiment, participants completed the tasks each day within windows of time, so participants in the same condition could have experienced somewhat longer or shorter delays based on exactly when they choose to complete the tasks. This regressions used each participants average delay intervals. The delay is coded in terms of hours for interpreting the regression coefficient. The distribution of the actual delay that each participant experienced was shown in Figure 5.

Table A1 presents the results of the regressions. Similar to the ANOVAs, there were no effects of delay for the causal strength measures before Trial 9 or after Trial 16, for the future prediction strength, and for future use strength. Similar to the ANOVAs participants who experienced longer delays had weaker trial-by-trial prediction strength, and the BF was stronger in the regression analysis. One difference was that there was a significant effect of delay on the frequency strength measure for the regression analysis but not the ANOVA analysis (p=.035) however, the BF was only 2.07.

Similar to the ANOVAs most of the regressions found an influence of dataset and none of them found an interaction between delay and dataset.

Figure 5: The distribution of average delays that participants actually experienced over 16 days in four delay conditions.



Note. The vertical dash lines indicate the ideal delay in each condition.

Table A1: Regression results for Causal Judgment measures.

	β	p	BF	η_p^2
Causal Strength (Before Trial 9)				
Delay	-0.003	0.417	0.37	<0.01
Dataset	-0.101	0.291	0.46	0.02
Delay:Dataset	-0.004	0.647	0.30	<0.01
Causal Strength (After Trial 16)				
Delay	-0.002	0.616	0.28	<0.01
Dataset	-0.241	0.010	5.52	0.07
Delay:Dataset	0.0002	0.981	0.25	<0.01
Future Prediction Strength				
Delay	-0.004	0.237	0.46	<0.01
Dataset	-0.261	0.001	27.3	0.07
Delay:Dataset	0.005	0.439	0.32	<0.01
Future Use Strength				
Delay	-0.003	0.525	0.30	<0.01
Dataset	-0.293	0.006	8.61	0.06
Delay:Dataset	0.004	0.661	0.27	<0.01
Frequency Strength				
Delay	-0.006	0.035	2.07	0.02
Dataset	-0.117	0.087	1.04	<0.01
Delay:Dataset	0.007	0.254	0.48	<0.01
Trial by Trial Prediction Strength				
Interval	-0.011	0.003	13.6	0.05
Dataset	-0.277	0.001	30.2	0.06
Interval:Dataset	0.009	0.229	0.45	<0.01

References

Aitken, M. R. F., Larkin, M. J. W., & Dickinson, A. (2001). Re-examination of the role of within-compound associations in the retrospective revaluation of causal judgements. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology*, *54B*(1), 27–51. doi: 10.1080/02724990042000029

Allan, L. G. (1980, March). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149. doi: 10.3758/BF03334492

Boakes, R. A., & Costa, D. S. J. (2014, October). Temporal contiguity in associative learning: Interference and decay from an historical perspective. *Journal of Experimental Psychology: Animal Learning and Cognition*, *40*(4), 381–400. doi: 10.1037/xan0000040

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018, December). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880–1910. doi: 10.1037/xlm0000548

Buehner, M. J. (2005, May). Contiguity and covariation in human causal inference. *Animal Learning & Behavior*, *33*(2), 230–238. doi: 10.3758/BF03196065

Buehner, M. J., & McGregor, S. (2006, November). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*(4), 353–378. doi: 10.1080/13546780500368965

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

Collins, D. J., & Shanks, D. R. (2002, October). Momentary and integrative response strategies in causal judgment. *Memory & Cognition*, *30*(7), 1138–1147. doi: 10.3758/BF03194331

Copelton, D. A., & Valle, G. (2009, August). "You don't need a prescription to go gluten-free": The scientific self-diagnosis of celiac disease. *Social Science & Medicine*, *69*(4), 623–631. doi: 10.1016/j.socscimed.2009.05.012

De Houwer, J., & Beckers, T. (2003, November). Secondary task difficulty modulates

forward blocking in human contingency learning. *The Quarterly Journal of Experimental Psychology Section B*, *56*(4b), 345–357. doi: 10.1080/02724990244000296

Dennis, M. J., & Ahn, W.-K. (2001, January). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition*, *29*(1), 152–164. doi: 10.3758/BF03195749

Fasano, A., Sapone, A., Zevallos, V., & Schuppan, D. (2015, May). Nonceliac Gluten Sensitivity. *Gastroenterology*, *148*(6), 1195–1204. doi: 10.1053/j.gastro.2014.12.049

Gallistel, C. R., Craig, A. R., & Shahan, T. A. (2019). Contingency, contiguity, and causality in conditioning: Applying information theory and Weber's Law to the assignment of credit problem. *Psychological Review*, *126*(5), 761–773. doi: 10.1037/rev0000163

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*(2), 289–344. doi: 10.1037/0033-295X.107.2.289

Gibson, P. R., Skodje, G. I., & Lundin, K. E. A. (2017). Non-coeliac gluten sensitivity. *Journal of Gastroenterology and Hepatology*, *32*(S1), 86–89. doi: 10.1111/jgh.13705

Gluck, M. A., & Thompson, R. F. (1987). Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review*, *94*(2), 176–191. doi: 10.1037/0033-295X.94.2.176

Goedert, K. M., Harsch, J., & Spellman, B. A. (2005, August). Discounting and Conditionalization: Dissociable Cognitive Processes in Human Causal Inference. *Psychological Science*, *16*(8), 590–595. doi: 10.1111/j.1467-9280.2005.01580.x

Griffiths, T. L., & Tenenbaum, J. B. (2005, December). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. doi: 10.1016/j.cogpsych.2005.05.004

Grossberg, S., & Schmajuk, N. A. (1989, January). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, *2*(2), 79–102. doi: 10.1016/0893-6080(89)90026-9

Hagmayer, Y., & Waldmann, M. R. (2002, October). How temporal assumptions influence causal judgments. *Memory & Cognition*, *30*(7), 1128–1137. doi: 10.3758/BF03194330

Hattori, M., & Oaksford, M. (2007). Adaptive Non-Interventional Heuristics for Covariation Detection in Causal Induction: Model Comparison and Rational Analysis. *Cognitive Science*, *31*(5), 765–814. doi: 10.1080/03640210701530755

Hogarth, R. M., & Einhorn, H. J. (1992, January). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. doi: 10.1016/0010-0285(92)90002-J

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1363–1386. doi: 10.1037/0278-7393.19.6.1363

Lagnado, D. A., & Speekenbrink, M. (2010, July). The Influence of Delays in Real-Time Causal Learning. *The Open Psychology Journal*, *3*(1). doi: 10.2174/1874350101003010184

Leslie, A. M., & Keeble, S. (1987, April). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288. doi: 10.1016/S0010-0277(87)80006-9

Liu, P.-P., & Luhmann, C. C. (2013, April). Evidence that a transient but cognitively demanding process underlies forward blocking. *Quarterly Journal of Experimental Psychology*, *66*(4), 744–766. doi: 10.1080/17470218.2012.717952

Logue, A. W. (1979). Taste aversion and the generality of the laws of learning. *Psychological Bulletin*, *86*(2), 276–296. doi: 10.1037/0033-2909.86.2.276

Marsh, J. K., & Ahn, W.-K. (2006, April). Order effects in contingency learning: The role of task complexity. *Memory & Cognition*, *34*(3), 568–576. doi: 10.3758/BF03193580

Michotte, A. (1963). *The perception of causality*. Oxford, England: Basic Books.

Mutter, S. A., Atchley, A. R., & Plumlee, L. M. (2012). Aging and retrospective revaluation of causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 102–117. doi: 10.1037/a0024851

Renner, K. E. (1964). Delay of reinforcement: A historical review. *Psychological Bulletin*, *61*(5), 341–361. doi: 10.1037/h0048335

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review*, *74*(1), 71–80. doi: 10.1037/h0024109

Rescorla, R. A., & Wagner, A. (1972, January). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory* (Vol. Vol. 2).

Revusky, S. (1971). The role of interference in association over a delay. In *Animal Memory* (W. K. Honig & P. H. R. James ed., pp. 155–213). New York, NY: Academic Press.

Schneiderman, N. (1966). Interstimulus interval function of the nictitating membrane response of the rabbit under delay versus trace conditioning. *Journal of Comparative and Physiological Psychology*, *62*(3), 397–402. doi: 10.1037/h0023946

Schneiderman, N., & Gormezano, I. (1964). Conditioning of the nictitating membrane of the rabbit as a function of CS-US interval. *Journal of Comparative and Physiological Psychology*, *57*(2), 188–195. doi: 10.1037/h0043419

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989, May). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology Section B*, *41*(2), 139–159. doi: 10.1080/14640748908401189

Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of Experimental Psychology*, *38*(2), 168–172. doi: 10.1037/h0055873

Smith, M. C., Coleman, S. R., & Gormezano, I. (1969). Classical conditioning of the rabbit's nictitating membrane response at backward, simultaneous, and forward CS-US intervals. *Journal of Comparative and Physiological Psychology*, *69*(2), 226–231. doi: 10.1037/h0028212

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020, July). Time and Singular Causation—A Computational Model. *Cognitive Science*, *44*(7). doi: 10.1111/cogs.12871

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA, US: The MIT Press.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies.* Lewiston, NY, US: Macmillan Press. doi: 10.5962/bhl.title.55072

Waldmann, M. R., & Walker, J. M. (2005, May). Competence and performance in causal learning. *Learning & Behavior*, *33*(2), 211–229. doi: 10.3758/BF03196064

Well, A. D., Boyce, S. J., Morris, R. K., Shinjo, M., & Chumbley, J. I. (1988). Prediction and judgment as indicators of sensitivity to covariation of continuous variables. *Memory & Cognition*, *16*(3), 271–280. doi: 10.3758/BF03197760

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J.

(2011, May). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(3), 291–298. doi: 10.1177/1745691611406923

Willett, C. L., & Rottman, B. M. (2020). Causal learning with two causes over weeks. *Proceedings of the 42nd annual conference of the cognitive science society*, 2007–2013.

Willett, C. L., & Rottman, B. M. (2021). The Accuracy of Causal Learning Over Long Timeframes: An Ecological Momentary Experiment Approach. *Cognitive Science*, *45*(7), e12985. doi: 10.1111/cogs.12985

Young, M. E., & Sutherland, S. (2009, August). The spatiotemporal distinctiveness of direct causation. *Psychonomic Bulletin & Review*, *16*(4), 729–735. doi: 10.3758/PBR.16.4.729