Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-023-02342-x

BRIEF REPORT

Causal learning with delays up to 21 hours

Yiwen Zhang[1] · Benjamin M. Rottman[1]

Accepted: 13 July 2023

Abstract
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. We found that learning was slowed with longer delays, but by the end of 16 days participants had learned the cause-effect relation in all four conditions, and they had learned the relation about equally well in all four conditions. This suggests that in real-world situations people may still be fairly accurate at inferring cause-effect relations with delays if they have enough experience. We also discuss ways that delays may interact with other real-world factors that could complicate learning.

Keywords Delay · Causal learning · Ecological momentary experiments

Introduction

In everyday life, causes can have an influence on their effects with considerable delays. For example, when assessing whether gluten has a negative impact on one's health, people need to connect the foods they ate and their symptoms hours or days apart (Fasano et al., 2015). However, a critical limitation of research on learning from experience (e.g., causal learning, correlation detection, reinforcement learning) is that studies have only investigated learning with delays on the order of seconds. The goal for the current study was to understand the role of hours long delays on the ability to accurately detect cause-effect relations.

Delays in conditioning and reinforcement learning

For decades it was believed that temporal contiguity is crucial to contingency learning in animal conditioning and reinforcement learning (Renner, 1964; Rescorla, 1967; Skinner, 1948). Many studies have found that longer delays impede learning. For instance, in classical trace conditioning, the rate of conditioning is inversely related to the intra-trial interval (Smith et al., 1969; Schneiderman and Gormezano, 1964; Schneiderman, 1966). In instrumental conditioning, animals have more difficulty acquiring a response when the reinforcements are delivered with longer delays (see Renner (1964); Boakes and Costa (2014) for reviews).

However, there are still open debates about the role of delay. For example, 'preparedness of learning' research has found an exception to the negative impact of delay; animals can learn with delays up to 24 h with food-related conditioned stimuli, likely an evolutionary adaptation to avoid foods that are poisonous (Logue, 1979). Additionally, most of the prior research focused on the time interval between the cue and outcome (intra-trial interval), but not the inter-trial interval. Gallistel and Gibbon (2000) proposed 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. In sum, there are still open questions as to when and why delays matter in animal conditioning (e.g., Boakes and Costa (2014), pp 395).

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. Even short delays destroy the perception of causal launching (Leslie and Keeble, 1987; Michotte,

The registration plans and data are available here: https://osf.io/qthme.
This research was funded by the National Science Foundation 1651330.

✉ Yiwen Zhang
yiwenzhang@pitt.edu

[1] Uinversity of Pittsburg, Pittsburgh, USA

Published online: 14 August 2023

1963; Young and Sutherland, 2009), and initial studies found that delays longer than 4 s reduce causal judgements of action-outcome relations in free-operant conditioning (Shanks et al., 1989).

However, subsequent studies showed that learning is not necessarily weakened by longer delays and instead is mediated by temporal assumptions (Buehner and McGregor, 2006; Buehner, 2005; Hagmayer and Waldmann, 2002). Buehner and colleagues argued that Shanks et al. (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 expected a long delay. These studies by Buehner and colleagues involved a paradigm in which there were a series of cause and effect events happening over continuous time. With longer delays it was more likely that a cause event might be followed by no effect event for a while or that another effect could occur due to a hidden background cause between the target cause-effect period. Knowledge that the relation involved a delay helped participants parse out which cause and effect events went together, termed the "attribution shift hypothesis."

Longer delays and current study

With the exception of the preparedness of learning research with animals, all the prior research has focused on seconds-long delays. The goal of the current study is to investigate how well people are able to learn cause-effect relations with delays on the order of hours. In prior human studies participants only had to remember the events occurring in the past few seconds, and memory was not considered of key importance, but with hours-long delays memory is crucial. For example, when learning the impact of eating gluten, one would need to keep track of the foods that they eat over many 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 to mimic real world causal learning, which we call "ecological momentary experiments." 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 and Rottman, 2021). People also incorrectly inferred "illusory correlations" when there was not a cause-effect relation roughly the same in spaced-out and rapidly presented conditions. Follow-up research focused on people's ability to learn about two causes and one effect in a long timeframe setting (Willett and Rottman, 2020). Similarly, Wimmer et al. (2018) tested reward-based learning when data were presented rapidly vs. spaced out; performance was not different between two conditions in an immediate test after learning, but performance was maintained better in the spaced condition when tested after 3 weeks (Wimmer et al., 2018).

Even though these three studies are more realistic than standard studies in that the trials were spaced out, for all three there was no delay between the cause and effect or action and feedback. Thus the findings possibly represent an overly optimistic picture of real-world learning. In the current study we tested the influence of delays up to 21 hours.

Aside from the role of temporal assumptions discussed above, there are two main reasons that learning may be slowed by delays (Boakes and Costa, 2014). One theory has to do with the number of intervening events; this will be covered in the general discussion. The other, especially relevant for this study, is the possibility of decay of the memory of the cause. If the leaner incorrectly remembers the cause as being present when it was absent or vice versa by the time that the effect occurs, the learning process would be noisy and therefore slowed down. Relatedly, even if the learner could accurately recall the cause after a long delay, they may not always spontaneously do so when experiencing the effect and therefore might not learn from each experience. There are two main accounts of causal learning (Perales and Shanks, 2007). Associative theories assume that people learn cause-effect relations by sequentially updating associative weights between the candidate causes and effects. "Rule-based" theories propose that people keep tallies of the four types of experiences (cause and effect present vs. absent), and then make causal judgments from these tallies. Delay could slow learning in both accounts due to noisy updating or failure to update tallies or weights. Our primary goal was to establish whether long delays impair human learning.

Methods

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 reported writing down data during the study and 1 participant who experienced a programming error.

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.

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. The 16 trials were pseudo-randomly ordered; the first and second halves had the same number of each of the A-D event types. The main reason for testing the two datasets was to distinguish learning from bias (e.g., a bias that participants on average believed that the medicine would be helpful or harmful). It also allowed us to test if learning was faster for positive or negative datasets.

Temporal delays

We manipulated the temporal delays within each trial; see Fig. 1. 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 later (min = 2, max = 7). In the 9-hour delay condition, participants did the cause task in the morning and the effect task in the evening around 9 hours later (min = 8, max = 15). In the 21-hour delay condition,

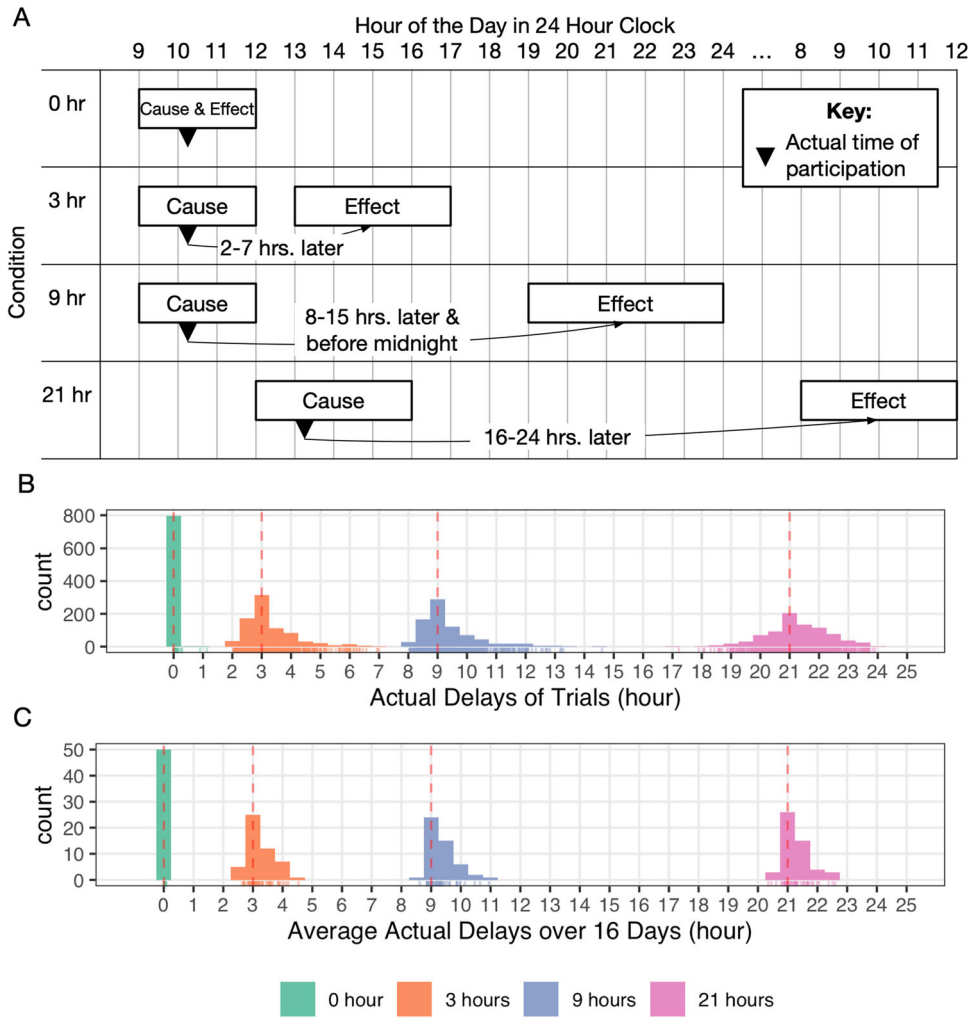

Fig. 1 The time windows for participation and histograms of actual delays

participants did the cause task in the afternoon and the effect task the next morning roughly 21 hours later (min = 16, max = 24).

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 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.

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.

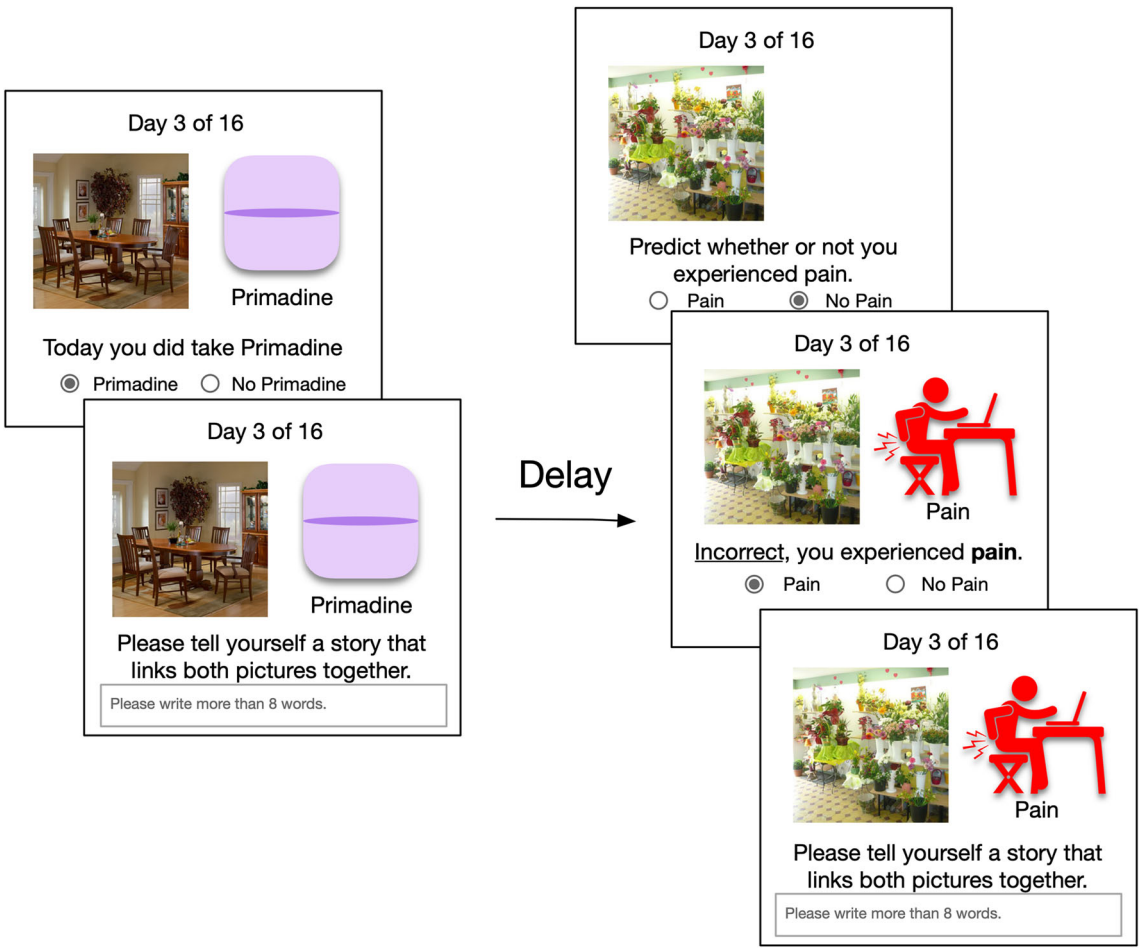

Fig. 2 Experimental flow of the cause task and the effect task

> 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 (Fig. 2). In the cause task, participants first saw a contextual image.[1] After they clicked the 'Continue' button, they were shown an icon and text of whether the cause was 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."

The effect task followed a similar procedure with a different contextual image (Fig. 2), except that before seeing whether the effect was present or absent, participants made a binary prediction about 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; the methods and results for these memory measures appear in the appendix.

[1] The images were taken from existing sources (Konkle et al., 2010; Robin and Olsen, 2019) as well as some of our own images and are available on our OSF registry. The contextual images were meant to represent a variety of different sorts of easy to identify categories (e.g., flower shop, dining room, beach). In the instruction, we told participants that "You will also see pictures of a scene. These pictures are supposed to represent places that you visited or things that you saw each day. Please try to remember this whole event - the picture and whether or not you took the medicine in the morning, as well as the picture and whether or not you had pain in the afternoon. To help improve your memories, try to tell yourself a story about the relation between the pictures. For example, suppose that you see a picture of a beach and an image of medicine in the morning. You could imagine taking medicine as you are packing up to go to the beach." The contextual images were randomly shuffled across all cause and effect events.

Measures

All the measures were scaled in a range of [-1,1] for analysis.

Prediction Strength During Learning

To have a measure that tracks learning over time, we computed "prediction strength during learning" from participants' binary predictions about the presence or absence of the effect: $p(\text{predicted effect} \mid \text{cause}) - p(\text{predicted effect} \mid \neg\text{cause})$. We calculated this for the first half of the learning trials (Trials 2-8),[2] and for the second half (Trials 9-16).

Causal strength

Participants made a "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 sort of causal strength question is the most common type of question in studies on causal learning and inference. This question was asked both halfway through the learning phase (before Trial 9) and in the testing session after learning. The remaining questions were only asked at the end of learning.

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). This measure was intended to be very similar to the predictions during learning, but assessed only at the end of the learning phase.

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. We have started asking this question in the current study as well as other related studies as an alternative to the causal strength measure. We

[2] An anonymous reviewer suggested that we also analyze the first half of the learning trials, which was not included in the preregistration. Trial 1 was excluded because on the first trial participants have learned nothing yet and therefore have no basis on which to make a prediction.

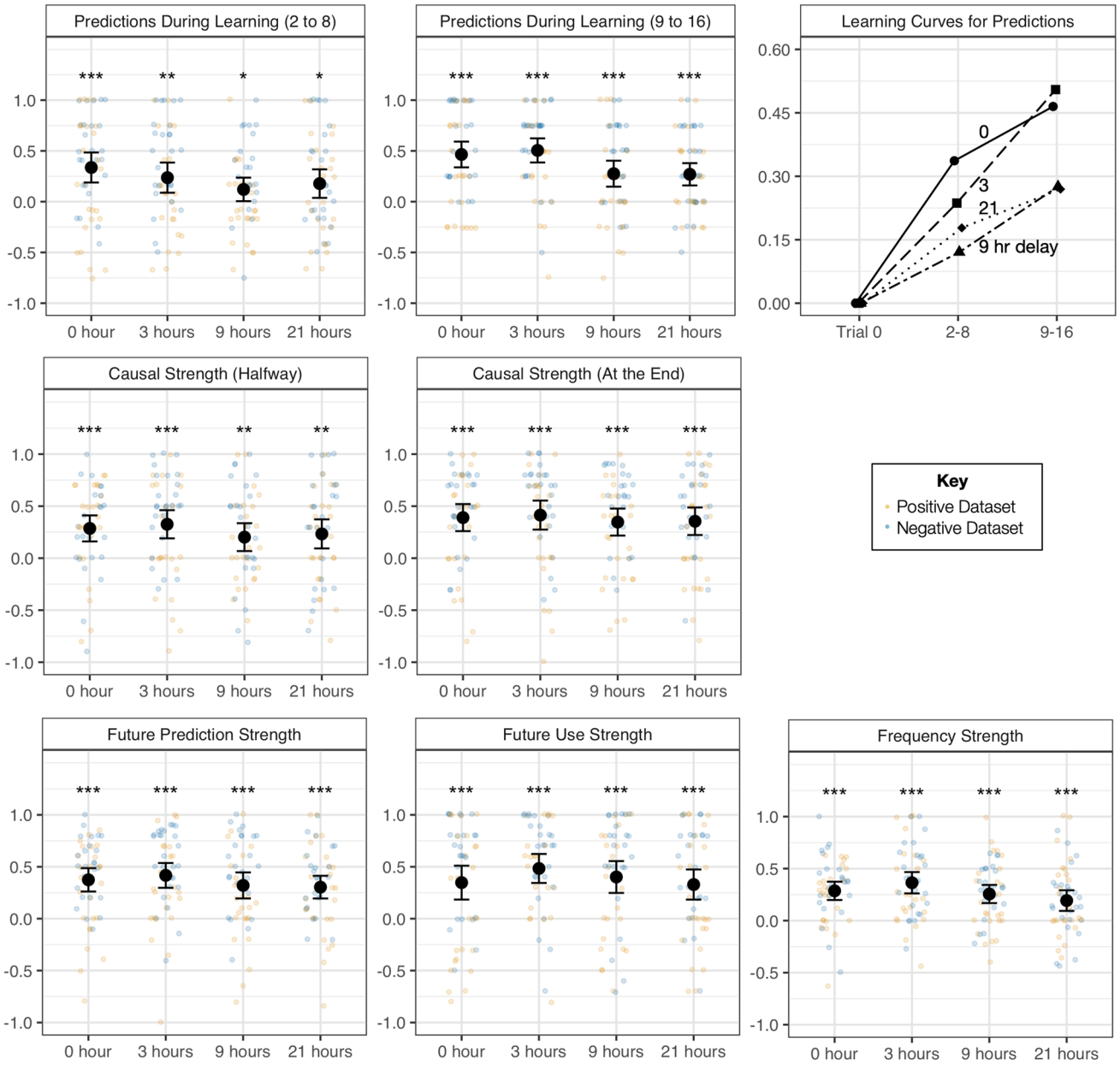

Fig. 3 Mean and 95% confidence interval of all measures separated by the four conditions

hypothesize that it gets at a similar idea as causal strength but is more behaviorally oriented.[3]

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; their frequency strength could not be calculated due to a division by zero problem, which can happen if some pairs of cells are judged as zero. Since "rule-based" theories of causal learning presume that people store tallies of these four quantities and use them to infer the strength of the relation, frequency strength provides an additional perspective into such theories.

[3] We also note that the wording of the causal strength measure ("improve pain" vs. "worsen pain") can be confusing for some participants. We believe that it is hard to word that measure more clearly, but feel that the future use strength gets at a similar idea in a different way.

Results

The analysis follows our pre-registered plan available at https://osf.io/qthme. All Bayes Factors (BFs) are presented such that numbers greater than 1 are evidence for the alternative, and less than 1 are evidence for the null. The BFs were calculated with the BayesFactor package (Morey et al., 2022) and we used the default priors.[4] For ease of interpretation, we inverse coded the judgements for the negative datasets so that they are positive.

Figure 3 shows all of the dependent measures; the Predictions During Learning are also plotted as learning curves in the top right panel. Participants' judgments were significantly above zero for all measures, at all timepoints, 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. Specifically, the *p*s were less than 0.05 for all 24 conditions, the BFs were larger than 8.60 for all but the Predictions During Learning Trials 2-8 in the 9 hours (BF = 1.15) and 21 hours (BF = 2.79) conditions, and the Cohen's d's ranged from 0.30 to 1.21.

We conducted two analyses to test the influence of delay. One compared the four delay conditions using an ANOVA. The other was a linear regression in which the predictor was the average delay for a given participant (see the distribution of the actual time intervals in Fig. 1) and rescaled the predictor so that 0 means no delay and 1 means an average delay of 21 hours. The ANOVAs and regressions tested for main effects of delay, dataset (positive vs. recoded negative) and their interaction. If learning becomes weaker with longer delays, there would be a main effect of delay.[5] Table 1 presents the results.

We first discuss the main effect of delay. There are three measures during the learning phase. For the Predictions During Learning for Trials 2-8 and the Causal Strength Halfway Through Learning there was not a significant influence of delay, and the BFs were in favor of the null to various degrees (0.05-0.63). In contrast, the Predictions During Learning measure for Trials 9-16 found weaker judgments with longer delays; this finding was significant according to the ANOVA (p =0.006, BF = 5.18) and regression (p =0.003, BF = 13.98).[6][7] We compared the 0 vs. 3, 3 vs. 9, and 9 vs. 21-hour delay conditions, while also controlling for the two datasets. 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, $\eta_p^2 = 0.07$.

We now focus on the main effect of delay in the judgments at the end of learning. There was no significant effect of delay for causal strength after learning, future use strength, or future prediction strength; the BFs for the ANOVAs were fairly strong in favor of the null (0.03-0.05) while those for the regression were less strong in favor of the null (0.27-0.46). In the frequency strength measure, the p-values were around the border of significant, and the BFs were fairly weak (one slightly in favor of the null and the other slightly in favor of the alternative).[8]

In sum, we found some evidence of slowed learning due to delay in the Predictions During Learning for Trials 9-16, but not for Trials 2-8 nor for Causal Strength Halfway Through Learning. With regards to the measures collected after the end of learning, only one of the four, Frequency Strength, yielded any hints of a remaining effect of delay.

We now move on to the main effect of dataset and the interaction of delay and dataset. Four out of the six ANOVAs also found very strong main effects of dataset. Because the negative condition was reverse coded, the main effect means that the recoded judgments in the negative condition are more positive/stronger than the judgments in the positive condition. Conceptually, participants gave stronger judgments when the medicine improved the symptom than when it worsened the symptom. None of the analyses showed an interaction between delay and dataset and all the BFs were in favor of the null.

[4] We present the results with the default setting of priors in this paper. We also tried different scale options for the default Cauchy prior and the results are available on osf. Overall, different prior scale options barely affect our interpretation of the results but when the scaling parameter becomes wider, the BF is a little bit more in favor of the null hypothesis.

[5] 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. One slight change from the registration is that we included dataset in the ANOVA; positive datasets are often judged more strongly than negative ones (Catena et al., 2004; Maldonado et al., 1999); if there is a difference then controlling for dataset (after reverse coding) increases power to detect a main effect of delay.

[6] These BFs were obtained with default Cauchy priors provided by the BayesFactor package in R. The default scale parameter is 1/2. When setting the scale parameter for the Cauchy priors to $\sqrt{2}/2$ (wide) and 1 (ultrawide), the BF decreased to 3.25 and 1.64 for the ANOVA results, and to 12.66 and 10.63 for the regression results respectively.

[7] In our registration we did not propose correcting alpha, despite having multiple dependent measures and two different tests. We feel that the dependent measures capture somewhat different things, and the two tests were run because we expected some sort of monotonic relation but not necessarily linear. We also feel that focusing on the p-value, BF, effect size, and the two tests provides the most compelling interpretation, rather than focusing just on the p-value. We acknowledge that readers may have differences of opinion about this.

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

Table 1 ANOVA and Regression Results for all measures

	ANOVA				Regression			
	F	p	BF	η_p^2	β	p	BF	η_p^2
Prediction During Learning (Trials 2 to 8)								
Delay	1.81	0.146	0.23	0.03	−0.12	0.148	0.63	0.01
Dataset	15.94	<0.001	>100	0.08	−0.32	0.001	35.30	0.07
Interaction	0.46	0.710	0.05	<0.01	0.13	0.435	0.31	<0.01
Predictions During Learning (Trials 9 to 16)								
Delay	4.28	0.006	5.18	0.06	−0.23	0.003	13.98	0.05
Dataset	12.12	0.001	45.13	0.06	−0.28	0.001	29.84	0.06
Interaction	0.77	0.510	0.13	0.01	0.18	0.230	0.45	<0.01
Causal Strength Halfway Through Learning								
Delay	0.62	0.603	0.05	<0.01	−0.07	0.417	0.37	<0.01
Dataset	3.98	0.048	1.04	0.02	−0.10	0.291	0.46	0.02
Interaction	1.13	0.340	0.19	0.02	−0.08	0.647	0.30	<0.01
Causal Strength at the End of Learning								
Delay	0.18	0.908	0.03	<0.01	−0.04	0.616	0.27	<0.01
Dataset	14.17	<0.001	>100	0.07	−0.24	0.010	5.51	0.07
Interaction	0.27	0.846	0.08	0.01	0.00	0.981	0.24	<0.01
Future Prediction Strength								
Delay	0.76	0.518	0.06	0.01	−0.09	0.237	0.46	<0.01
Dataset	14.75	<0.001	>100	0.07	−0.26	0.001	27.27	0.07
Interaction	0.57	0.635	0.10	0.01	0.11	0.439	0.32	<0.01
Future Use Strength								
Delay	0.85	0.467	0.06	0.01	−0.06	0.525	0.30	<0.01
Dataset	12.57	0.001	55.71	0.06	−0.29	0.006	8.61	0.06
Interaction	0.36	0.783	0.08	<0.01	0.08	0.661	0.27	<0.01
Frequency Strength								
Delay	2.24	0.085	0.41	0.03	−0.13	0.035	2.07	0.02
Dataset	1.65	0.200	0.34	<0.01	−0.12	0.087	1.04	<0.01
Interaction	0.46	0.709	0.09	<0.01	0.14	0.254	0.48	<0.01

The analyses for the memory questions are available in the appendix; briefly, there were no significant differences across the four conditions, and the BFs were in the direction of the null.

Discussion

This is the first experiment to test human learning with hours-long delays. There were three key findings. First, longer delays slowed down learning as measured by the predictions during learning. Second, participants were able to eventually learn the cause-effect relation even with delays up to 21 hours. Third, by the end of the 16 days of learning, most of the evidence was in favor of a null effect of delay suggesting that participants overcame the long delays.

The finding of the weaker predictions for Trials 9-16 with longer delays could be explained with a slower learning rate. Indeed, according to standard reinforcement learning models (e.g., Rescorla and Wagner, 1972), a difference in learning rate between conditions can appear as a larger difference in cue weights mid-way through learning compared to right at the beginning of learning or farther along. At the same time, a curious aspect of the finding of slowed learning is that the typical impact of delay in associative learning is a lower asymptote (or that the relation cannot be learned at all), rather than slowed learning with a similar asymptote (Boakes and Costa, 2014; Sutton and Barto, 1990). The weaker predictions for Trials 9-16 with longer delays was also not reflected in the causal strength judgments, revealing some inconsistency in the findings.

One ancillary finding was that participants learned negative better than positive relations. It is likely that participants

thought that it was more plausible that a medicine prevented disease symptoms (the negative condition) rather than caused symptoms as side effects (the positive condition). Note that this pattern is the opposite of the finding in associative learning that positive relations are learned faster than negative (Wagner and Rescorla, 1972), which has also been found in human causal learning studies (Catena et al., 2004; Maldonado et al., 1999).

Multiple Theories and Multiple Measures

We collected multiple dependent measures for two reasons. First, the precise way of asking questions about causality can lead to different findings (Collins and Shanks, 2006; Matute et al., 2002); asking multiple questions could provide more certainty about the consistency of results. Second, some of the questions are more relevant to certain theories than others. Dissociating different theories of causal learning is notoriously challenging (Shanks, 2007), and was not the primary goal of the current study. However, certain patterns of results could have revealed that some theories are more likely than others. We discuss the episodic memories in the Appendix and the others below.

The frequency strength measure is most closely aligned with "rule-based" theories that assume that people store tallies of the four types of events (Perales and Shanks, 2007). If people implement an associative learning process, it may be possible to derive tallies from associative weights, but doing so is not straightforward and not part of standard associative theory. We found some evidence that the frequency strength judgments were worse with longer delays, potentially suggesting that tally-based learning processes may be used less with long delays. However, because this evidence is weak, this hypothesis is speculative.

In contrast, other measures, such as future prediction strength, future use strength, and causal strength could be computed easily from tallies or from associative weights, so do not discriminate between theories. The predictions during learning are required by associative and reinforcement learning theories, and though they are not required by other theories, could easily be implemented by any theory of causal learning. Thus, the finding that learning is slowed with delay does not implicate a particular theory.

Limitations and Open Questions

Though this research provides an important step towards understanding learning in more real-world settings with memory demands, there are still many open questions. First, one of the main theories about delay is that there are more intervening events with longer delays, and it is not the delay that leads to worse learning but the number of intervening events (Boakes and Costa, 2014; Lagnado et al., 2010; Revusky, 1971). The fact that our study was conducted through a smartphone app, and there was only a single candidate cause, means that there were no other "relevant" alternative intervening causes; presumably participants could easily filter out everything happening in their life outside this app. This allows for increased focus on the single target cause and simplifies the credit assignment problem. But in real-world situations, there will almost always be other intervening alternative causes, which could exacerbate the impact of delay.

Second, another important theory about human causal learning is that learning is impaired when the delay deviates from the learner's expectations (Buehner and McGregor, 2006). To mitigate this, we picked cover stories for which we thought that short and long delays were plausible; if anything zero delay is implausible. Furthermore, unlike prior studies that used a continuous time paradigm in which participants could expect the outcome to occur earlier or later than when it occurred, this was not possible in our study because they knew that the effect would be shown at one preset time. Thus, we think that expectations probably had a fairly minor role in this study, but it is possible that they had some role.

Third, the current study did not examine the impact of delay variability; cause-effect relations with the same average delay but larger variability in delay tend to be rated as weaker (Greville and Buehner, 2010, 2016). In our task participants knew that the effect would be conveyed to them at a specific time each day, so even if there was variability in when they performed the task, this is quite different and simpler than situations in which a learner must figure out 1) whether a cause will produce an effect, 2) if so when it will occur, and 3) ensure that the effect did not occur because of an alternative cause. Testing the influence of delay variability with hours-long delays is a natural extension of the current research. Studies with delays on the order of seconds have also found that people use the variability and correlations among delays for inferring the causal structure among three variables (Bramley et al., 2018) and for judging which of two candidate causes actually produced an effect (Stephan et al., 2020). Testing these phenomena with long delays are also important future directions.

Fourth, given that the task could happen while participants were doing other things, if learning was at floor it could be explained merely due to a lack of processing. We wanted to ensure that participants were paying attention and encoding the stimuli, not just clicking through the task. Thus, we asked participants to write short stories about each learning episode to encourage attention. We also had participants make predictions during learning; doing so is believed to have a neutral or slightly beneficial impact on learning (Well et al., 1988), perhaps due to increased attention or for reasons similar to retrieval practice (Roediger and Karpicke, 2006;

Rowland, 2014). A potential concern is that these procedures may have led to artificially increased salience or attention, perhaps leading to an overly optimistic picture of learning with delays. Though possible, we think there are reasons not to be too concerned. The stories were still quite short and likely took only 10-20 s to write. In comparison, many real-world events 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. With regards to the predictions, reinforcement-learning theories assume that people spontaneously make predictions, in which case making explicit predictions may not have much of an impact nor be artificial.

Conclusions

This research makes an important empirical contribution to the field of human causal learning, and learning more generally, showing that learning is both slowed by delays, but that eventually people can effectively learn even despite long delays. 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 and with enough trials. Still, it is important to try to test more complex and realistic learning situations, which may interact with delay.

Appendix Episodic Memories

In addition to associative learning and keeping track of tallies of the four event types, another way that people might learn cause-effect relations is through storing episodic memories of the cause-effect events, and sampling from these events when making a prediction or causal judgment. Episodic memory has not received much attention within the causal learning literature, however, some theories of reinforcement learning posit that people utilize episodic memories in addition to cue weights (Bornstein et al., 2017; Bornstein and Norman, 2017; Gershman and Daw, 2017). We hypothesized that longer delays between the cause and effect may lead to greater decay of the cause memory by the time that the effect occurs, and therefore they may have worse episodic memory for the cause-effect pairs.

Methods

There were eight probes during the memory task, and for each probe there were two parts. In the recognition task, participants were shown two pairs of contextual images. One pair was comprised a true pair of contextual images that occurred during the cause and effect events. The other pair, a lure, comprised a pair of two contextual images from a cause event and an effect event but from different nonsuccessive days. We randomly picked 8 learning trials be the veridical target trials, and the images from the other trials were used for the lures. For the recognition task, participants were asked to identify the pair of images that they saw in the same trial (Fig. 4 Part 1). After making this choice, the participant received feedback about whether their choice was correct or incorrect, and were shown the two contextual images that actually appeared during the same cause-effect pair.

For the episode memory task, while these two cause-effect contextual images were being shown participants were asked to recall whether the cause happened or not, and whether the effect happened or not, on the day that the contextual image occurred. As participants provided their responses for the cause and effect, the cause-present or cause-absent, and effect-present or effect-absent images appeared, similar to during the learning phase. When they were ready they submitted their answer for both the cause and effect, and did not receive feedback (Fig. 4 Part 2).

The participants completed 8 trials in the memory task - the images for the other 8 episodes were used for the lures. For the recognition memory task, each trial was coded as correct or incorrect, and the average of the 8 trials calculated; chance was 0.5. For the episode memory task, each trial was coded as correct if participants choose the correct state of both the cause and effect. The 8 trials were averaged together, and chance was 0.25.

Results

First, we compared the memory accuracy with the chance level (Table 2). Single-group t-tests revealed that the recognition accuracy was higher than the chance level in the 3 and 9-hour delay conditions, and not significantly different from the chance level in the 0 and 21-hour delay conditions. Episode memory accuracy was higher than the chance level in all conditions. Note, however, that for the episode memory, if participants learned a positive relation, they could just tend to say that both the cause and effect were absent, or both present, and have accuracy above 0.25 without actually having specific episodic memories for these cause-effect events. Given that the recognition memory for the contextual images was only slightly above chance, this possibility seems likely.

The most important analysis of the memories was to test for an influence of delay. Analogous to the two analyses in the main results in the manuscript, we ran both ANOVAs and linear regressions, for both the recognition memory and episode memory data. For the recognition memory, both the ANOVA ($F = 1.19$, $p = 0.31$, $BF = 0.11$) and the regression ($b = -0.02$, $p = 0.21$, $BF = 0.32$) were not significant with BFs in the direction of the null. And for episode memory, both the ANOVA ($F = 1.51$, $p = 0.21$, $BF = 0.16$) and the regression (b

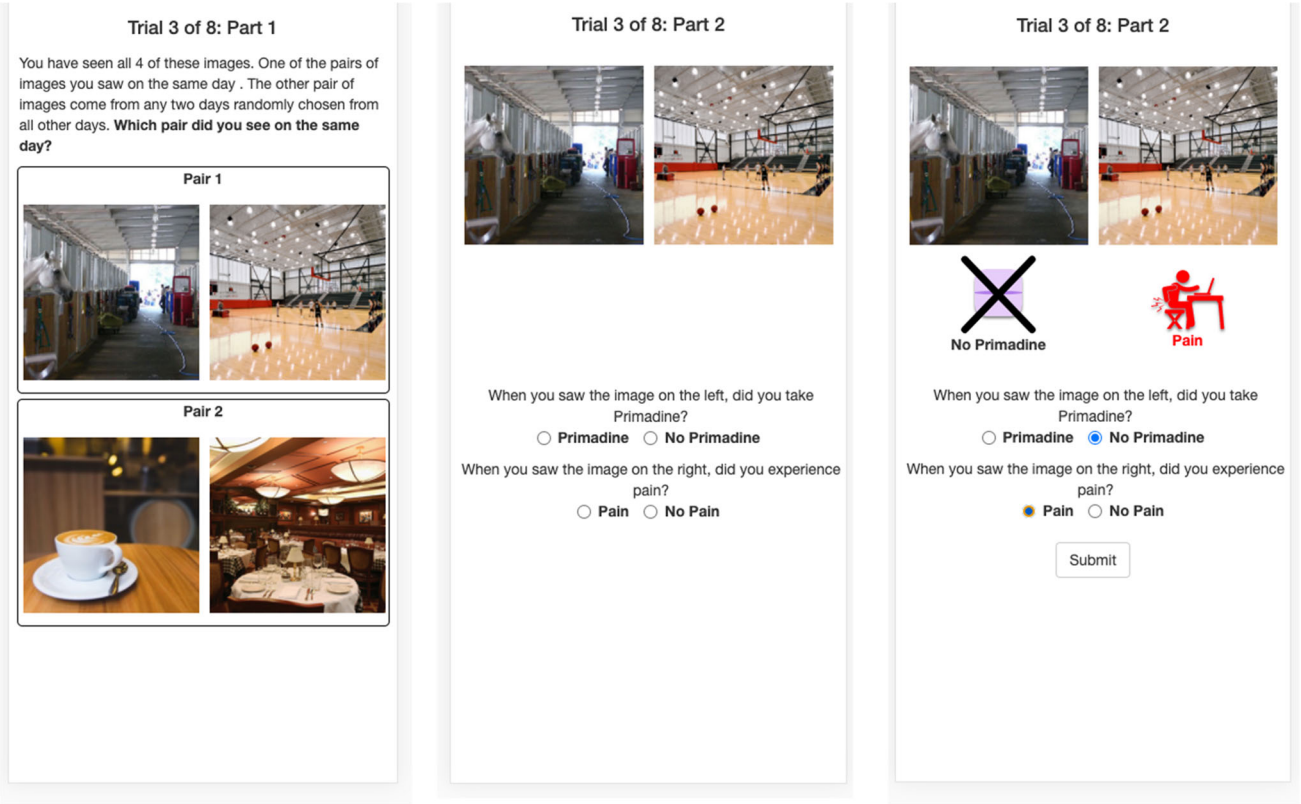

Fig. 4 An example probe in the memory task

=0.01, p =0.97, BF =0.15) were not significant with BFs in the direction of the null.

Discussion

There is no evidence of an influence of delay on the episodic memories. This is consistent with the lack of delay on the other summary measures at the end of learning. It is also possible that no influence of delay can be detected if these memories are near floor. Recognition memory was only slightly above chance. And though the episode memories were significantly above chance, this could be due to post-hoc reconstruction without veridical memories of the cause and effect events.

Table 2 Comparing memory accuracy against chance levels

	Mean [95% CI]	t	p	BF	cohen's d
Recognition Memory (chance =0.5)					
0 h	0.55 [0.49, 0.60]	1.57	0.124	0.48	0.22
3 h	0.56 [0.51, 0.61]	2.37	0.022	1.94	0.34
9 h	0.57 [0.52, 0.62]	3.07	0.004	9.34	0.44
21 h	0.51 [0.46, 0.56]	0.308	0.759	0.16	0.04
Episodic Memory (chance =0.25)					
0 h	0.35 [0.30, 0.40]	4.03	<0.001	>100	0.57
3 h	0.43 [0.38, 0.48]	6.67	<0.001	>100	0.94
9 h	0.39 [0.34, 0.44]	5.61	<0.001	>100	0.81
21 h	0.38 [0.33, 0.44]	5.03	<0.001	>100	0.71

In our previous studies in which the cause and effect were presented simultaneously on the same screen with only contextual image there was also fairly weak evidence of episodic memories in the long timeframe; there was stronger evidence of episodic memories including primacy and recency effects when the study was conducted in rapid setting in which the learning events were back-to-back (Willett et al., underreview).

It is still possible that people sample from episodic memory but only form bindings of the cause and effect events and do not bind them with the contextual images. If so, this account would be fairly similar to keeping tallies. It is also possible that there are more subtle effects, such as recency effects that guide predictions during learning, which may be stronger with shorter delays.

Acknowledgements This work was supported by NSF 1651330. The authors thank 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.

Declarations

Conflicts of interest We have no known conflicts of interest to disclose.

References

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

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

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications, 8*, 15958.

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature neuroscience, 20*, 997–1003.

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

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

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

Catena, A., Perales, J. C., & Maldonado, A. (2004). Judgment frequency effects in generative and preventative causal learning. *Psicologica: International Journal of Methodology and Experimental Psychology*, *25*, 67–85.

Collins, D. J., & Shanks, D. R. (2006). Conformity to the power pc theory of causal induction depends on the type of probe question. *Quarterly Journal of Experimental Psychology, 59*, 225–232.

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

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

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology, 68*, 101–128.

Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General, 139*, 756.

Greville, W. J., & Buehner, M. J. (2016). Temporal predictability enhances judgements of causality in elemental causal induction from both observation and intervention. *Quarterly Journal of Experimental Psychology, 69*, 678–697.

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

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological science, 21*, 1551–1556.

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

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

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

Maldonado, A., Catena, A., Cándido, A., & García, I. (1999). The belief revision model: Asymmetrical effects of noncontingency on human covariation learning. *Animal Learning & Behavior, 27*, 168–180.

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

Michotte, A. (1963). *The Perception of Causality. The Perception of Causality*. Oxford, England: Basic Books.

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2022). BayesFactor: Computation of Bayes Factors for Common Designs.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic bulletin & review, 14*, 577–596.

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

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

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

Revusky, S. (1971). The role of interference in association over a delay. In *Animal Memory* (pp. 155–213). New York, NY: Academic Press. (W. k. honig & p. h. r. james ed.).

Robin, J., & Olsen, R. K. (2019). Scenes facilitate associative memory and integration. *Learning & Memory, 26*, 252–261.

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science, 17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432–1463. https://doi.org/10.1037/a0037559

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*, 397–402. https://doi.org/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*, 188–195. https://doi.org/10.1037/h0043419

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology, 60*, 291–309. https://doi.org/10.1080/14640748908401189

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

Skinner, B. F. (1948). "Superstition" in the pigeon. *Journal of Experimental Psychology, 38*, 168–172. https://doi.org/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*, 226–231. https://doi.org/10.1037/h0028212

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

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (pp. 497–537). Cambridge, MA, US: The MIT Press.

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

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*, 271–280. https://doi.org/10.3758/BF03197760

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). 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*, 291–298. https://doi.org/10.1177/1745691611406923

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

Willett, C., & Rottman, B. M., (2020). Causal learning with two causes over weeks. In S. Denison, M. Mack, & Y. Xu, & B. C. Armstrong (Eds.) Proceedings of the 42st Annual Conference of the Cognitive Science Society. Austin TX: Cognitive Science Society.

Willett, C. L., Zhang, Y., & Rottman, B. M. (under review). Causal learning with two causes over weeks, .

Wimmer, G. E., Li, J. K., Gorgolewski, K. J., & Poldrack, R. A. (2018). Reward Learning over Weeks Versus Minutes Increases the Neural Representation of Value in the Human Brain. *The Journal of Neuroscience, 38*, 7649–7666. https://doi.org/10.1523/JNEUROSCI.0075-18.2018

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