# The Accuracy of Causal Learning over 24 Days

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

## Abstract

Humans often rely on past experiences stored in long-term memory to predict the outcome of an event. In traditional lab-based experiments (e.g., causal learning, probability learning, etc.), these observations are compressed into a successive series of learning trials. The rapid nature of this paradigm means that completing the task relies on working memory. In contrast, real-world events are typically spread out over longer periods of time, and therefore long-term memory must be used. We conducted a 24 day smartphone study to assess how well people can learn causal relationships in extended timeframes. Surprisingly, we found few differences in causal learning when subjects observed events in a traditional rapid series of 24 trials as opposed to one trial per day for 24 days. Specifically, subjects were able to detect causality for generative and preventive datasets and also exhibited illusory correlations in both the short-term and long-term designs. We discuss theoretical implications of this work.

**Keywords:** causal learning; probability learning; illusory correlation; long-term memory; smartphone

## Introduction

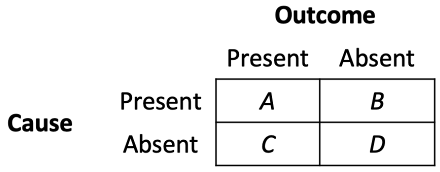
Every day we use our experiences to make inferences. For example, is your new medication improving an ailment or causing a negative side-effect? Does meditating have a positive impact on your mental health? If we can accurately predict the outcomes of our experiences and actions, we can use this information to behave adaptively in the world.

Trial-by-trial learning paradigms, in which cue-outcome pairs are presented to subjects sequentially, are used extensively to study learning across many different fields including causal learning, probability learning, fear learning, stereotype formation, associative learning with non-human animals, and others. The trial-by-trial paradigm is supposed to simulate an important aspect of the world: most of our experiences occur sequentially over time, rather than in a summarized form. Typically the 'inter-trial-interval', the time between trials, is a couple seconds. However, we contend that there are few real-world learning situations that involve experiencing repeated cue-outcome pairs separated by seconds, perhaps with a few exceptions (e.g., flipping through records rather than first-hand experiences).

The goal of the current study is to compare trial-by-trial learning in the normal rapid format vs. trial-by-trial learning in which the trials are spaced out once per day. Day-by-day learning simulates many natural processes (e.g., does a medicine have an influence on a health outcome, does exercising have an influence on sleep, etc.). Importantly, whereas working memory is believed to support learning in short timeframes, long-term memory must take over when learning occurs over many days. In the current study we investigated how effectively people are able to learn cue-outcome relations across multiple days.

### Trial-by-Trial Causal Learning

Prior research has evaluated how people detect causation from data shown over a successive series of trials. In a typical experiment, participants observe data in which the putative cause and the outcome are either present or absent. This information can be organized into a 2x2 table where each cell *A-D* represents the number of times that the cause/outcome combination occurs for a particular dataset (see Figure 1). Most often, participants are shown the data rapidly, for example two or three seconds per trial. After observing the entire dataset, subjects judge the degree to which the cause influences the outcome.



| | | **Outcome** | |
|---|---|---|---|
| | | Present | Absent |
| **Cause** | Present | *A* | *B* |
| | Absent | *C* | *D* |

Figure 1: A 2x2 table depicting the four possible types of data in a traditional binary design.

One normative model of causation is the $\Delta P$ rule, a measure of contingency that suggests an optimal way to infer causation is by comparing the probability of the outcome in the presence of the cause and the probability of the outcome in the absence of the cause: $\Delta P = A/(A+B) - C/(C+D)$. When $\Delta P$ is positive, the causal relationship is generative. When $\Delta P$ is negative, the causal relationship is preventive.

Although prior research suggests that people are able to discriminate generative vs. preventive causation (Shaklee & Mims, 1982), individuals sometimes exhibit biases in causal reasoning. One such bias, "illusory correlation" or "illusory causation", occurs when people inaccurately infer causation when no causal relationship exists.

An "A-cell bias" is when individuals believe that a causal relation exists merely because of a high number of A-cell trials (e.g., Kao & Wasserman, 1993). In the A-cell bias condition in Table 1, even though there is zero relation between the cue and outcome (the outcome occurs with a chance of .625 regardless of whether the cue is present or absent, so $\Delta P = 0$), people tend to infer that they are positively correlated. An "outcome density bias" is when people incorrectly assign causation to a dataset in which the overall probability of the outcome is high (Table 1), even though the probability of the outcome is the same (.75) whether the cause is present or absent, so Δ$P$ = 0 (e.g., Jenkins & Ward, 1965).

Table 1: Cell Frequencies for the 4 Datasets

| Dataset | $A$ | $B$ | $C$ | $D$ | Δ$P$ |
|---|---|---|---|---|---|
| Generative | 9 | 3 | 3 | 9 | 0.5 |
| Preventive | 3 | 9 | 9 | 3 | -0.5 |
| Outcome-Density | 9 | 3 | 9 | 3 | 0 |
| A-cell | 10 | 6 | 5 | 3 | 0 |

## Causal Learning and Memory

Many causal learning experiments use rapidly successive trial-by-trial paradigms. In the real world, however, you would not experience each data point in rapid succession. This raises a number of challenges for long term memory. For example, imagine learning whether going to yoga improves your mood; some days you do yoga and other days you do not. After a few weeks, would you be able to remember the days you did or did not do yoga? Could you remember your mood on those days? How might your memories for these events impact your ability to detect causation? Would you be more susceptible to biases such as illusory correlations? Currently, there is no research on how well people can learn causal relations over long timespans.

One basis for making hypotheses about causal learning in long timeframes is research on short timeframes that has increased working memory (WM) demands. Studies have found stronger illusory correlations in a rapid trial-by-trial paradigm (higher WM demands) than in a "summary" paradigm (lower WM demands) in which all the trials are presented simultaneously (Kao & Wasserman, 1993). Adding a distractor task on top of the trial-by-trial paradigm leads to less accurate judgments (Shaklee & Mims, 1982), and older adults with lower WM have less accurate causal learning (Mutter & Pliske, 1996). If causal learning is worse when WM is taxed, we expected learning to get even worse when long-term memory must be used to assess causation. Still, people are often able to navigate the world successfully, suggesting a reasonable causal-learning ability when relying on long-term memories to make inferences. This raises the question: how well can we learn causal relations across many days?

## Summary of Current Study

In the current study, we investigated the implications of learning a cause-effect relationship quickly from a rapid sequence of trials vs. learning the same relationship over an extended period of time – one trial per day for 24 days. We investigated how subjects learned about four causal relations using different datasets: generative, preventative, 'outcome-density', and 'A-cell' (Table 1).

The motivation for studying the generative and preventive datasets was  to determine whether or not participants were capable of detecting a causal relationship or if learning is hampered when the experiences occur spread out in time. Because memories might be noisier in the long-term condition, we predicted that participants' judgments might be closer to zero, implying a weaker causal relationship.

For the A-cell and outcome density datasets, we wanted to assess the effect of long-term memory on illusory correlations. Prior research mainly found exaggerated illusory correlations with increased WM demand, so one hypothesis was that illusory correlations would be exaggerated in the long-term condition. Another hypothesis was that, if memories of the experiences are weaker in the long timeframe condition, then the judgments might actually be closer to zero – more accurate.

# Methods

## Participants

There were 476 participants. The main requirements were owning a smartphone and intending to complete the entire study; however, we mainly targeted college students to have a similar sample to most other causal learning studies and since they frequently use smartphones.  Participants were paid $30 if they successfully completed the entire study.

Our goal was to have around 400 participants, 100 for each of the 4 datasets in the long timeframe condition. The large number was used because the four datasets need to be analyzed separately, and to have power to detect small effects. The final data analyses included 409 participants after dropping 13 participants who admitted to writing down data during the study, 1 who was not trying during the task, 39 due to a programming error, and 14 who skipped too many days of the long timeframe task.

## Datasets

Participants learned about five datasets: four short-timeframe (generative, preventative, A-cell, and outcome density) and one long-timeframe (one of the four from the short-timeframe

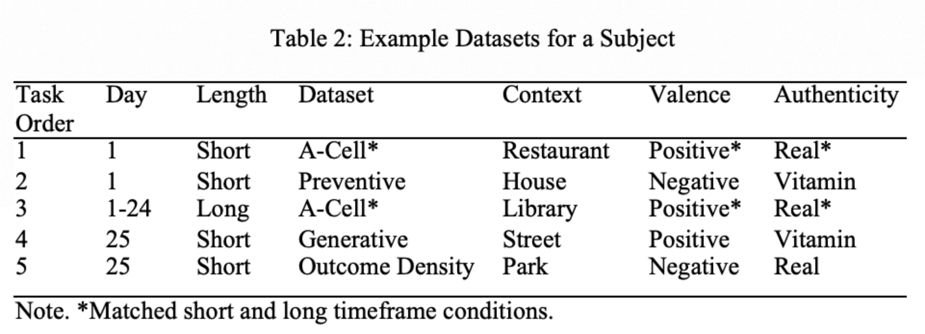


Table 2: Example Datasets for a Subject

| Task Order | Day | Length | Dataset | Context | Valence | Authenticity |
|---|---|---|---|---|---|---|
| 1 | 1 | Short | A-Cell* | Restaurant | Positive* | Real* |
| 2 | 1 | Short | Preventive | House | Negative | Vitamin |
| 3 | 1-24 | Long | A-Cell* | Library | Positive* | Real* |
| 4 | 25 | Short | Generative | Street | Positive | Vitamin |
| 5 | 25 | Short | Outcome Density | Park | Negative | Real |

Note. *Matched short and long timeframe conditions.

condition). This design allowed for a within-subjects comparison of one of the four datasets across the long vs. short conditions (see Table 2 for an example). By having subjects learn all four datasets in the short timeframe condition, it also reduces the likelihood that subjects were aware that one of the short timeframe datasets was the same as the long timeframe dataset. Each dataset consisted of 24 trials ordered randomly. The two illusory correlation datasets were previously used by Kao and Wasserman (1993).

## Procedure

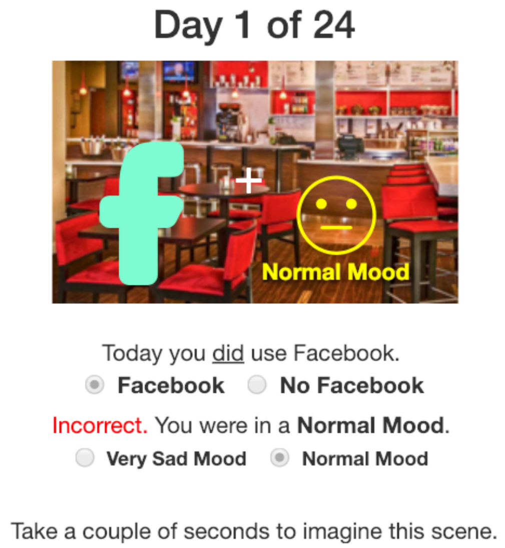
Participants completed the entire study on their own smartphones by logging into our website created with our PsychCloud.org framework. The procedure for the short-term and long-term tasks were identical, except that subjects observed one trial per day in the long timeframe condition, and they did trials back-to-back in the short timeframe condition. On Day 1 of the study, participants completed two short-term tasks and began Day 1 of the long-term task.

On Days 2 – 24, participants received automated text-message reminders at 10am, 3pm, and 8pm to complete their daily trial for the long-term task and stopped receiving reminders if they had already participated that day. They returned to the lab on Day 25 to complete the remaining short-term tasks and receive payment. The order of the short-term tasks was randomized so that participants completed the short version of the long-term task either on Day 1 or on Day 25 - before or after the long-term task.

**Within a Trial** Each task consisted of 24 trials in which participants were told whether or not the putative cause was present or absent. A number of procedures were taken to facilitate encoding, including asking subjects to verify the state of the cause and effect (rather than just observe them), and to spend extra time to look each image. Each trial proceeded as described in the following example, which uses the 'Facebook' cover story – other cover stories are explained below. In the Facebook cover story, subjects were asked to judge whether using Facebook during their lunch break improves or worsens or has no influence on their mood, based on the hypothetical dataset.

At the beginning of each trial, subjects were shown a contextual image. These images allowed us to ask a number of episodic memory questions that are not analyzed in this report. In the Facebook cover story, they saw an image from the inside of a restaurant and were told "This is the scene from your lunch break." After three seconds, an icon and text were superimposed over the contextual image to show the presence or absence of the cause (e.g., whether they used or did not use Facebook during their lunch break). They pressed a radio button to confirm the state of the cause and could not move on until selecting the correct button (e.g., Facebook vs. No Facebook). Next, they pressed a radio button to predict the effect as present or absent (e.g., Very Sad Mood vs. Normal Mood). They received text feedback for whether their prediction was correct or incorrect and an icon representing the effect was superimposed on the image. After clicking the correct radio button to verify the state of the effect, subjects were instructed to "Take a couple of seconds to imagine this scene", which was displayed for an additional four seconds.

At the end of a trial in the short timeframe condition, subjects were permitted to move on to the next trial. In the long timeframe condition, subjects were told that their task was over and to come back to the website the following day. Once a trial was over, the website did not allow subjects to see the data for that trial or prior trials, not even by clicking the back button on their web browser.



Day 1 of 24

Today you did use Facebook.

● Facebook ○ No Facebook

Incorrect. You were in a **Normal Mood**.

○ Very Sad Mood ● Normal Mood

Take a couple of seconds to imagine this scene.

Figure 2. Screenshot of the end of a trial.

After Trial 24 (either immediately afterwards for the short timeframe condition, or on Day 25 in the long timeframe condition), participants judged the strength of the causal relationship. First, they answered whether the cause (Facebook) "improves or worsens or has no influence" on the effect (mood). If participants said the cause had no influence, they were assigned a causal judgment of 0. If they responded "improve" or "worsen", they answered "How strongly does [the cause] [improve/worsen] [the effect]?" on a scale of 1 (*very weak*) to 10 (*very strong*), which produced a scale from -10 to +10. In this report we only discuss the judgments after Trial 24, though subjects also made similar judgments before Trial 9 and before Trial 17.

In addition to the causal strength judgments, participants also made a number of other judgments, for example, memories of the number of experienced cells of types A, B, C, and D (before Trials 5, 13, 21, and after Trial 24), as well as a number of judgments about the memories for the contextual images after Trial 24. These measures will not be analyzed in the current report due to space.

**Cover Stories** Since subjects learned about five cause-effect relations, we created the following five authentic '*contexts*', randomly assigned to the five tasks so that each was viewed as a separate learning task: the relation between using Facebook during lunch in a *restaurant* and mood, eating a healthy dinner in a friend's *house* and having an upset stomach, using notecards to study in a *library* and grades on a daily quiz, biking to work on *city streets* and productivity at work, and bringing your dog on a walk in a *park* and stress. The five stories were chosen so that it would be plausible for the cause to either improve or worsen the outcome; the influence of prior beliefs will be analyzed in other reports.

Because this study is the first to use a long timeframe paradigm, is unlikely to be replicated, and is focused on external validity, we conducted two manipulations of the cover stories. Specifically, we manipulated the "authenticity" and "valence" of the cover stories. If subjects in the long timeframe condition exhibited very poor learning, we wanted to rule out some potential explanations and to know how to best design future studies. Although we will explain the manipulations here, they are not of primary importance and will not be analyzed in this report.

First, though it is typical in causal learning studies to use entirely novel and abstract cover stories to minimize the influence of prior beliefs, we worried that abstract stimuli could be hard to remember in a long timeframe condition.[1] For this reason, we manipulated the 'authenticity' of the cover stories. The 'authentic' cover stories were the five stories mentioned previously. In the 'novel' cover stories, we used the same effects but replaced the causes with a hypothetical vitamin that a subject took on some days but not others (e.g., does the vitamin have an influence on mood, upset stomach, etc.). The matched short-term and long-term datasets were assigned to different contexts but were matched on authenticity. Of the four short timeframe conditions, two were assigned to 'novel' vitamin cover stories and two were assigned to authentic cover stories (Table 2).

Second, we manipulated the 'valence' of the effect; whether the presence of the effect is good or bad.[2] The absence of the effect was always described as normal (e.g., normal mood, normal grade on a quiz, etc.). The presence of the effect was described as either very good or very bad (e.g., very happy or very sad; very good grade or very bad grade, etc.). For participants in the negative valence condition, we reverse coded their causal strength judgments, so positive causal strength means "improved" for the positive valence condition and "worsened" for the negative valence condition. The matched short-term and long-term datasets were assigned the same valence. Of the four short-term conditions, two had positive and two had negative valence (see Table 2). Authenticity and valence are not analyzed due to space.

**Participation** Before starting the experiment, participants were told that if they missed more than three days in the long timeframe task, the study would be terminated and that they would not be paid. 462 (97%) participants successfully completed the study. On any given day, 83% of subjects participated before the 3pm reminder, 96% before the 8pm reminder, and 99% by midnight. If a subject missed one, two, or three days, the subsequent days were automatically pushed back the appropriate number of days.

The causal strength judgments and other measures for the long timeframe task occurred during the second in-lab testing session. We worked hard to have subjects come back to the lab for the second in-lab testing session on Day 25, one day after the last trial in the long timeframe condition. Of the 409 subjects in the final analyses, 83% returned to the lab on Day 25. If they skipped one day of the long timeframe task, sometimes this session occurred on the same day as their 24th trial (13%). If the session had to be moved, sometimes it occurred two (3%) or three (1%) days after the last trial. Overall, the protocol was followed with high fidelity.

# Results

## Causal Strength Judgments

In this paper, we only analyzed data from the matched short-term and long-term conditions. We analyzed the generative ($N = 98$), preventive ($N = 102$), A-cell ($N = 105$), and outcome density ($N = 104$) conditions separately.

Average causal strength judgments are presented in Figure 3. Significance values above each column indicate whether the value was significantly different from zero. The significance value above the horizontal lines indicates whether the judgments in the short and long-term conditions were significantly different from each other. We calculated Bayes Factors (*BF*) for each t-test, where a $BF > 1$ is support for the alternative hypothesis and a $BF < 1$ is support for the null. Often *BF*s > 10 (or < 1/10) are considered "strong" evidence for the alternative (or null), *BF*s > 30 or < 1/30 are considered "very strong" and *BF*s >100 or < 1/100 are considered "extreme" (e.g., Lee & Wagenmakers, 2013).

**Generative and Preventive Conditions** First, we wanted to assess whether participants were capable of detecting causation in the generative and preventive conditions. For the generative dataset, causal judgments were significantly different from zero in both the short-term condition, $t(97)$ =

[1] For example, we suspect that in short learning tasks using novel stimuli, subjects might use other cues such as the position of stimuli on the screen rather than the semantic meanings of the cues. Such alternative methods of learning might be less salient in the long timeframe condition. Instead, we thought that semantically meaningful cause-effect relations might be easier to remember and also have higher external validity.

[2] Most studies on causal learning use cues that are either present or absent. Presence/absence of the cause and the effect is theoretically important in some theories of causal learning (e.g., Cheng, 1997). Further, the definition of the cells as A-D only makes sense with cues that are present/absent (not "high"/"low" or "2"/"1", etc.; see Figure 1). In order to stick close to prior studies and to be able to study the A-cell bias, we used present/absent cues. However, one consequence of using presence/absence is that most outcomes have an implicit valence of being good or bad. For example, many prior studies have used outcomes like the presence/absence of a headache (bad) or of a flower blooming (good). We did not want to arbitrarily use outcomes of one particular valence, or to confound valence with cover story. Furthermore, valence can influence the strength of illusory correlations (Mullen & Johnson, 1990). For all these reasons, we counterbalanced the valence of the cover story.

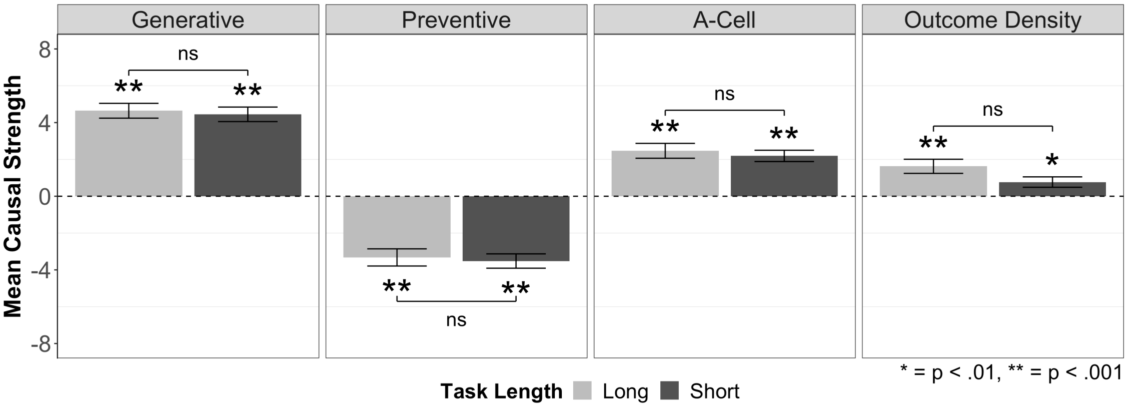


Figure 3: Average causal strength judgments after reviewing 24 trials. Error bars indicate standard error.

11.27, $p < .001$, $d = 1.14$, $BF = 2.13 * 10^{16}$, and the long-term condition, $t(97) = 11.53$, $p < .001$, $d = 1.17$, $BF = 7.37 * 10^{16}$. For the preventive dataset, judgments were less than zero in both the short-term, $t(101) = -9.03$, $p < .001$, $d = -0.89$, $BF = 5.72 * 10^{11}$, and long-term condition, $t(101) = -7.13$, $p < .001$, $d = -0.71$, $BF = 6.05 * 10^{7}$.

We predicted that for both the generative and preventive datasets, causal judgments would be closer to zero in the long-term condition because participants' memories would be noisier. However, paired t-tests revealed no significant differences between judgments in the short-term and long-term conditions for either the generative, $t(97) = -0.37$, $p = .707$, $d = -0.04$, $BF = 0.12$, or preventive datasets, $t(101) = -0.33$, $p = .741$, $d = 0.03$, $BF = 0.12$. Thus, participants were just as capable of detecting causation in the short and long timeframe conditions.

**Illusory Correlation Conditions** In the outcome-density and A-cell datasets, an optimal causal judgment would be zero. In line with our predictions, we found significant illusory correlations for both datasets**.** For the A-cell dataset, causal judgments were significantly greater than zero in both the short-term, $t(104) = 7.13$, $p < .001$, $d = 0.70$, $BF = 6.75 * 10^{7}$, and long-term, $t(104) = 6.11$, $p < .001$, $d = 0.60$, $BF = 6.36 * 10^{5}$, conditions. We found similar results for the outcome-density dataset; judgments were also positive and significantly different from zero in the short-term, $t(103) = 2.73$, $p = .008$, $d = 0.27$, $BF = 3.60$, and long-term, $t(103) = 4.23$, $p < .001$, $d = 0.41$, $BF = 341.33$, conditions.

We hypothesized that the illusory correlations could be either exacerbated or diminished in the long timeframe condition. However, there were no differences between causal judgments in the short and long-term conditions for the A-cell dataset, $t(105) = -0.67$, $p = .500$, $d = 0.07$, $BF = 0.13$. Illusory correlations appeared slightly stronger in the long-term condition for the outcome-density bias dataset, but this trend only approached significance, $t(104) = -1.87$, $p = .065$, $BF = 0.45$, with a small effect size of $d = 0.18$. These results suggest that illusory correlations in traditional trial-by-trial experiments are similar to what we observe in a long timeframe task.

## Predictive Strength

Another way to measure learning, aside from causal strength judgments, is through subjects' predictions of whether the outcome was present or absent each day. To ensure that participants had observed enough experiences to make predictions, we analyzed the predictions from Trials 13 – 24.

We transformed participants' predictions into a measure of causal strength by subtracting the probability that they predicted that the outcome would be present given the absence of the cause from the probability that the outcome would be present given the presence of the cause. This measure of "predictive strength" is conceptually similar to $\Delta P$. These results are displayed in Figure 4.

**Generative and Preventive** We found very similar results using subjects' predictions to assess learning as from their causal strength judgments. In the generative condition, predictive strength was significantly greater than zero for both the short-term, $t(97) = 11.58$, $p < .001$, $d = 1.17$, $BF = 9.01 * 10^{16}$, and long-term conditions, $t(97) = 12.47$, $p < .001$, $d = 1.26$, $BF = 6.32 * 10^{18}$. In the preventive condition, predictive strength was significantly less than zero in both the short-term, $t(101) = -11.87$, $p < .001$, $d = -1.18$, $BF = 6.77 * 10^{17}$, and long-term conditions, $t(101) = -9.38$, $p < .001$, $d = -0.93$, $BF = 3.12 * 10^{12}$. We found no difference in predictive strength between the short-term and long-term conditions for either the generative, $t(97) = -0.36$, $p = .718$, $d = -0.04$, $BF = 0.12$, or preventive, $t(101) = -0.49$, $p = .623$, $d = -0.05$, $BF = 0.12$, datasets. In sum, participants learned to accurately predict the effect, to the same extent, in both conditions.

**Illusory Correlation Conditions** In the A-cell bias condition, we found a similar pattern of results to the strength judgments. Subjects did infer an illusory correlation; they were more likely to predict the effect as present when the cause was present in both the short-term, $t(104) = 3.66$, $p < .001$, $d = 0.36$, $BF = 51.08$, and long-term condition, $t(104) = 3.66$, $p < .001$, $d = 0.36$, $BF = 50.64$. Furthermore, we found no difference between predictions in the short-term vs. long-term conditions, $t(104) = -0.66$, $p = .512$, $d = 0.06$, $BF = 0.13$.

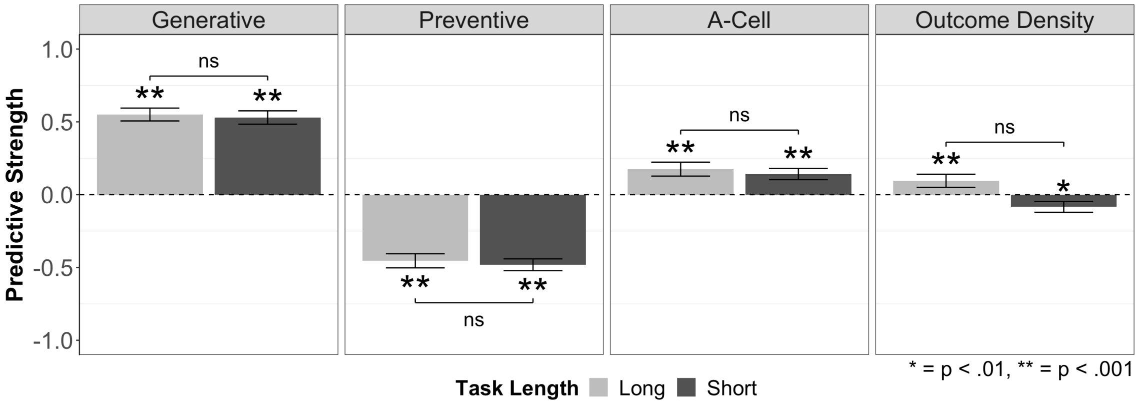


Figure 4: $\Delta$ P calculated from predictions about the effect given the cause. Error bars indicate standard error

In the outcome-density condition, the predictions were significantly negative in the short-term condition, $t(103) = -2.24$, $p = .028$, $d = -0.22$, $BF = 1.17$. However, they were significantly positive in the long-term, $t(103) = 2.13$, $p = .036$, $d = 0.21$, $BF = 0.94$, and the difference was statistically significant, $t(103) = -3.60$, $p < .001$, $d = -0.35$, $BF = 42.20$.

This difference was only marginally significant for the causal strength analyses, and the causal strength judgments for the short timeframe were significantly positive, not negative. Because this is the only difference between the two conditions, and it was only found for predictive strength (not the causal strength judgments) in the outcome density condition (not the other illusory correlation condition), we do not want to over-interpret it.

## Discussion

We sought to evaluate the external validity of traditional trial-by-trial causal learning experiments by comparing trial-by-trial learning when presented rapidly vs. one trial per day for 24 days. Presumably the former relies on working memory, whereas the latter requires long term memory. Our findings suggest that people are capable of learning generative and preventive causal relationships and also exhibit illusory correlations when learning causal relations over 24 days. Critically, we found few differences between the short-term and long-term tasks, and in fact most of the Bayes factors were roughly 8 to 1 in favor of the null.

From a practical perspective, this research provides an optimistic perspective on the validity of the trial-by-trial paradigm as a simulation of causal learning that occurs in the real world across longer periods of time. Assessing the external validity of this paradigm is important given that it has been used in hundreds of published studies on causal learning, and many thousands of studies when including studies of probability learning and other related topics.

From a theoretical perspective, we find it striking that there are so few differences in learning across the short and long timeframe condition. We intentionally used large samples to have the power to detect small effects. The robust learning in the long timeframe condition is surprising considering that participants completed the long-term trials outside of the lab and likely participated with many distractors and interruptions, comparable to everyday causal learning. Still, we hypothesized that the learning in the long timeframe condition would be plagued by considerably worse learning due to noisy memories. The fact that we found few differences raises a number of questions.

One question has to do with how learning occurs (e.g., Bornstein et al., 2017). Are subjects recording individual episodic memories and using them for causal learning? Or are they merely encoding them as generic events of the four cell types? Or are they using a process more similar to reinforcement learning in which an estimate of the strength of the relation between the cause and outcome gets updated as new evidence is experienced? Some of these questions can be addressed with our contextual image memory questions.

Another question is how well long-term memory can support other types of learning. It is possible that a single cause-effect relation is simple enough for long-term memory to robustly support learning, but that long-term memory might not be able to support more complex cause-effect relations (e.g., with multiple causes or long delays). We are actively studying such questions.

This research also has potential implications for whether learning and memory processes are fundamentally the same for shorter vs. longer timeframes. In associative learning, there is a debate about “timescale independence or invariance” (Gallistel & Gibbon, 2000), in which learning phenomena tend to replicate if the sequence is stretched or compressed. In memory, there are debates about the similarities and differences in short vs. long-term memory (e.g., Cowan, 2008) and whether memories across short and long timespans can be modeled with the same forgetting curves (e.g., Wixted & Ebbesen, 1991). Perhaps researchers invested in these debates may be able to use these results.

More generally, we believe that the current research provides an important step towards generalizing current learning paradigms to more real-world settings. The current findings are optimistic in terms of how well the paradigm generalizes; however, future research may also reveal areas in which standard learning paradigms generalize poorly.

## Acknowledgments

This work was supported by NSF 1651330. We would like to thank all of the research assistants who helped with data collection, including Aleks Brown, Beatrice Langer, Caitlin Viele Haggerty, Chris Shon, Elise Faut, Gabriela Cuadro, Joanna Ye, Julia Gillow, Melinda Rosen, Michael Datz, Minbae Le, Priya Chandrasekaran, Rachel Hopkins.

## References

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.

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological review*, *104*(2), 367-405.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Processes in brain research, 169,* 323-338.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review, 107*(2), 289-344.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79*(1), 1-17.

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.

Lee, M. D., & Wagenmakers, E-J. (2013). Bayesian cognitive modeling: A practical course. Cambridge University Press.

Mullen, B. and Johnson, C. (1990), Distinctiveness-based illusory correlations and stereotyping: A meta-analytic integration. *British Journal of Social Psychology, 29*, 11–28. doi: 10.1111/j.2044-8309.1990.tb00883.x

Mutter, S. A., & Pliske, R. M. (1996). Judging event covariation: Effects of age and memory demand. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *51*(2), 70-80.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(3), 208-224.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological science*, *2*(6), 409-415.