The Accuracy of Causal Learning over 24 Days

by

Ciara Willett

B.A., St. Mary's College of Maryland, 2014

M.S., Seton Hall University, 2017

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

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Ciara Willett

It was defended on

June 19, 2019

and approved by

Christian Schunn, Professor, Dept. of Psychology

Marc Coutanche, Assistant Professor, Dept. of Psychology

Thesis Advisor/Dissertation Director: Benjamin Rottman, Associate Professor, Dept. of Psychology

The Accuracy of Causal Learning over 24 Days

Ciara Louise Willett, M.S.

University of Pittsburgh, 2020

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.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Every day we use our experiences to guide our actions. For example, determining whether or not a new medication is improving an ailment (or causing a negative side-effect) could influence whether a patient decides to continue to use the medication. Or, determining whether meditating has a positive impact on one's mental health could influence their decision to continue to meditate. If we can accurately detect the relations between our experiences and their associated outcomes, we can then predict the outcomes of future actions and use this information to behave adaptively in the world.

The goal of the current study is to compare trial-by-trial learning in a paradigm in which the trials are presented fairly quickly (often just a few seconds per trial), with learning in which the trials are spaced out once per day. Whereas working memory is believed to support learning in short timeframes, long-term memory must take over when learning occurs over many days. The question then becomes how effectively people are able to learn cue-outcome relations across multiple days. In the current study we investigated trial-by-trial learning in a long timeframe by adapting a standard causal learning paradigm.

In the rest of the introduction, we first talk about the pervasiveness of trial-by-trial learning paradigms across many areas of psychology, and the need for research on trial-by-trial learning over longer durations of time. We then talk about prior research on how people learn causal relations and the role of working memory in accurately learning the strength of causal relations.

1.1 The Timing of Trial-by-Trial Learning

Trial-by-trial learning is an extremely common learning paradigm in which participants are presented with one cue (or many cues) and an outcome in a series of trials to learn the relation(s) between the cue(s) and outcome. The trial-by-trial paradigm simulates how individuals learn through trial and error while experiencing a temporal stream of events. Originally used in the behaviorist tradition, trial-by-trial learning is used pervasively across many fields including causal learning (e.g., Spellman, 1996; Waldmann, 2000), correlation detection (e.g., Jenkins & Ward, 1965; Kao & Wasserman, 1993), reinforcement learning (e.g., Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Delgado, Nystrom, Fissell, Noll, & Fiez, 2000), category learning (Kruschke, 1992; Nosofsky, 1986), fear learning (e.g., LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; Schiller et al., 2010), stereotype formation (Hamilton & Gifford, 1976; LePelley et al., 2010), and many other sub-fields.

One common aspect of the trial-by-trial paradigm is that the data is presented rapidly. Typically, each trial lasts a couple of seconds with only a couple of seconds between trials. 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., an admissions counselor flipping through records of students and looking at relations between variables such as grades and SAT, or other situations with previously-compiled records). Instead, most of our experiences that involve learning cue-outcome associations occur over considerably longer periods of time. For example, stereotypes are not learned rapidly on the order of seconds, but through experiences with in-group and out-group members that are most likely spaced out over days, weeks, months, or longer. Similarly, learning about potential food allergies or the effectiveness of a medication, common cover stories in the field of causal learning, is based on

experiences that are spaced out over days and weeks. In the current study, we compare rapid trial-by-trial learning with learning that occurs one trial per day. Day-by-day learning simulates many natural processes that often occur or do not occur once on a given day (e.g., does a medicine that can be taken once per day have an influence on a health outcome, does exercising on some days have an influence on sleep, etc.).

Only a few studies that have investigated learning over longer periods of time by modifying certain aspects of the learning and testing paradigms. For example, one study tested the effect of inter-trial interval length on depressive realism in causal learning (Msetfi, Murphy, Simpson, & Kornbrot, 2005), but the maximal inter-trial interval was still only 15 seconds. Many studies within the animal learning literature have manipulated the inter-trial interval, but the longest ITIs that we know of were still less than 30 minutes (e.g., Carranza-Jasso, Urcelay, Nieto, & Sánchez-Carrasco, 2014; Holland & Morell, 1996; Mustaca, Gabelli, Papini, & Balsam, 1991; Sheffeld, 1950; Stanley, 1952). Animal conditioning studies of this nature are also difficult to translate to many human paradigms because a longer ITI is typically represented as a longer period in which the cue and outcome are both absent. In human learning studies, often the cue and outcome are explicitly presented as present or absent, such that each trial is presented for the same length of time and a longer ITI is simply a longer time between learning (cf., Msetfi et al., 2005). Other studies have used training over multiple days, but during each day the training involved many trials with short inter-trial intervals (e.g., de Wit et al., 2018; Frey, Mata, & Hertwig, 2015; Tricomi, Balleine, & O'Doherty, 2009; Wunderlich, Dayan, & Dolan, 2012; Wimmer, Li, Gorgolewski, & Poldrack, 2018). In sum, we know of no trial-by-trial learning studies in which the trials are spaced out considerably over time.

Another set of literature that is relevant to the current study is research on the spacing effect (also called the 'lag effect' or the 'distributed practice effect') in memory. Prior research suggests that distributed practice enhances performance in verbal recall (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), skill learning (Donovan & Radosevich, 1999), and categorical induction (Vlach, Sandhofer, & Kornell, 2008) tasks. A typical spacing paradigm involves two practice sessions and a third test session, and the length of time between the study sessions and between the second study session and the testing session are varied. Cepeda et al.'s (2006) meta-review of the verbal recall literature found significant improvements when the study sessions were farther apart in time. Subsequent research has found that for longer delays before test, the optimal inter-study-interval increases. Though there are small decrements if the inter-study-interval is too long, in general "the penalty for a too-short gap is far greater than the penalty for a too-long gap" (Cepeda et al., 2009). This literature is relevant because it has investigated inter-study-intervals of various times from very short delays to multiple days.

However, there are several differences between the tasks used in this literature and our paradigm that make it hard to make specific predictions about the influence of spacing in a trial-by-trial causal learning task. First, spaced learning in this literature (e.g., Cepeda et al., 2006) typically involves two (or a few) training sessions, and within each training session the subject learns about many items. In contrast, in our spaced causal learning task there are 24 spaced out learning opportunities with a single item. Second, paired associate verbal recall tasks are purely memory tasks, whereas causal learning and other probability and reinforcement learning tasks involve *inferring* the statistical relationship between two cues. A more conceptually similar task is that of categorical induction (e.g., learning to categorize paintings of different artists), which also exhibits spacing effects (Metcalfe & Xu, 2016; Vlach et al., 2008). In categorical induction

studies, however, spaced intervals are on the matter of seconds as opposed to days. Overall, though the distributed practice literature is different from our study in many ways, it appears to support a prediction of superior learning in the long-term task as opposed to short-term task in the current experiment.

In summary, we believe that many important real-world learning situations involve learning in which the experiences are spaced out roughly one day apart, or at least often hours apart. However, we know of no studies that have investigated this type of learning with substantial delays.

1.2 Trial-by-Trial Causal Learning

The current study is a causal learning study in which participants learn the statistical relationship between a cause and an outcome, both of which can be either present or absent. Stimuli of this nature are typically conceived of in 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). 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.

When learning the strength of the relation between a cue and outcome, there are two basic questions. First, are people able to detect a true statistical relation between the cue and outcome?

Second, are people able to detect the true absence of a statistical relation? There are a number of 'normative' statistical models used as benchmarks against which human judgments can be compared (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005). In the current paper we use the simplest normative model, the ΔP rule (Allan, 1980), which compares the probability of the outcome in the presence of the cause and the probability of the outcome in the absence of the cause: $\Delta P = p(o|c) - p(o|{\sim}c) = A/(A+B) - C/(C+D)$. When ΔP is positive, the causal relationship is generative such that the cause produces the outcome. When ΔP is negative, the causal relationship is preventive such that the cause inhibits the outcome. If ΔP is equal to 0, the relationship is non-causal.

Although prior research suggests that people are able to adequately detect some important aspects of causation such as discriminating between generative and preventive causal relationships (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. In this paper, we study two types of illusory correlations, the A-cell bias and the outcome-density bias.

An A-cell bias is said to occur when individuals believe that a causal relation exists merely because of a high number of A-cell trials (e.g., Allan & Jenkins, 1980; Kao & Wasserman, 1993; Blanco, Matute, & Vadillo, 2013). 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. Similarly, an outcome density bias is said to occur 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 $\Delta P = 0$

(e.g., Buehner, Cheng, & Clifford, 2003; Jenkins & Ward, 1965). In this research, we studied the four datasets in Table 1 to see whether people's ability to appropriately learn the relation between the cue and outcome differed in short vs. long timeframes.

Table 1 Frequencies for Datasets in Current Study

Dataset	A	B	C	D	p(o=1|c=1)	p(o=1|c=0)	ΔP
Generative	9	3	3	9	.75	.25	0.5
Preventive	3	9	9	3	.25	.75	-0.5
Outcome-Density	9	3	9	3	.75	.75	0
A-cell	10	6	5	3	.625	.625	0

Note. The frequencies for the A-cell and outcome-density datasets (the illusory correlation datasets) are the same as those used by Kao and Wasserman (1993).

1.3 Causal Learning and Memory

How might learning be influenced when the experiences are spaced out over longer periods of time, such as once per day, as opposed to massed together and separated? Theoretically, whereas short-term memory must be used in typical studies, long-term memory must be used to keep track of the relation between the cue and outcome when spaced out over many days. However, keeping track of the relation is likely to be challenging 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?

Models of causal learning make different predictions about the role of memory in causal judgments. Rule-based accounts of causality (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Hattori & Oaksford, 2007), assume that individuals use their memories for the events to judge causality. Thus, accurate memories for the number of observations in each of the four cells of the contingency table are necessary to make accurate judgments of the strength of the relation.

In contrast, reinforcement-learning and associative accounts of causal learning, such as the Rescorla-Wagner Model (1972), do not require accurate memories of the events. Instead, associative accounts suggest that individuals update their belief about the strength of the relationship between the cue and outcome when presented with new evidence. All that needs to be remembered from one point of time to the next is a single value, the associative strength. Thus, according to reinforcement-learning and associative accounts, people should be able to make fairly accurate estimates of causal strength even with fairly minimal memory resources, and without accurate memories of the experienced events. The current experiment is not meant to arbitrate between rule-based and associative accounts of causal learning (see Waldmann, 2000), but instead to understand how people learn causal relations when the learning is mediated through short-term vs. long-term memory due to the sort vs. long-term nature of the task.

One basis for making hypotheses about causal learning in long timeframes is research on learning over short timeframes with 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 to participants simultaneously, somewhat similar to the summary of frequencies in Figure 1 (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 abilities exhibit less accurate

causal learning (Mutter & Pliske, 1996). If causal learning is worse when WM is taxed, one hypothesis is that learning might get even worse when long-term memory must be used to assess causation instead of short-term memory. If the short vs. long timeframe task reflects the same differences between lower vs. higher WM demands, then the judgments of causal strength would be closer to zero for the generative and preventative datasets in the long timeframe condition, and they would exhibit more 'illusory correlation' in the A-cell bias and the outcome-density conditions in the long timeframe condition.

Another potential hypothesis is that memories will be noisier in the long timeframe condition due to greater opportunity for memory decay. Specifically, if noise is injected into the remembered tallies of the four types of events *A*-*D*, assuming that the noise is equally distributed across the four cells, the more noise, the closer the causal strength judgment would be to zero. Therefore, this noise-based account makes the same prediction as the WM analogy for generative and preventative datasets, that the causal strength judgment will be closer to zero in the long timeframe condition. Furthermore, noisier memories in the long timeframe condition would also predict that the judgments are closer to zero (i.e., weaker illusory correlation) in the A-cell and the outcome-density datasets in the long-term condition. This prediction is actually the opposite of the findings from studies that have manipulated working memory demand.

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? If people are able to learn cause-effect relations fairly well over long time periods, this could suggest a couple of different possibilities. First, they might simply learn the associative strength and forget the experienced

events. Or they might remember a distribution of the four type of events, but not store episodic memories of each individual event over the 24 days.

1.4 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' (see Table 1). There are a variety of potential predictions. Research on spaced vs. massed learning could be interpreted to predict better learning in the long timeframe (though this study does not involve delayed recall). Research on working memory predicts worse performance in the long timeframe condition if it is analogous to increased WM demand. A memory decay perspective predicts noisier judgments in the long timeframe condition, which would lead to less accurate judgments for the generative and preventative datasets but more accurate judgments for the two zero contingency datasets. More generally, it is crucial to know whether people are able to accurately learn causal relations in long timeframes.

2.0 Methods

2.1 Participants

There were 476 participants (mean age = 21 years, 97% under 30 years old). 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 admitted to not trying during the task, 39 due to a programming error, and 14 who skipped too many days of the long timeframe task.

2.2 Datasets and Design

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 condition). This design allowed for a within-subjects comparison of one of the four datasets across the long vs. short conditions (see Table 1 for the cell frequencies in each dataset and Table 2 for an example of the five tasks). 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). Participants were randomly assigned to observe one of the four datasets, with the order of the matched short-version randomized to appear before or after the corresponding long-term task. For the set of two tasks preceding and following the long-term task, one was contingent (generative or preventive) and one was non-contingent (outcome density or A-cell).

Table 2 Example Order of Tasks 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. * indicates a task in which the dataset, cover story valence, and cover story authenticity were matched, but the length of the task was either short or long.

2.3 Procedure

2.3.1 Overall 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.

2.3.2 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 (see Figure 2).

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.

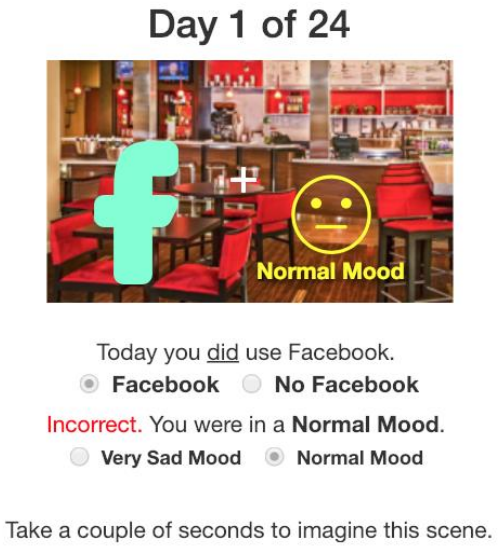

Figure 2 Screenshot of the end of a trial in which participants observe the data for at least 4 seconds before the task is completed.

2.3.3 Dependent Variables

2.3.3.1 Causal Strength

Participants judged the strength of the causal relation at three times. In the short timeframe condition this happened between Trials 8-9, 16-17, and immediately after Trial 24. In the long timeframe condition, this happened at the beginning of Days 9 and 17, and on Day 25.

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.

2.3.3.2 Frequency Judgments

Participants provided memories of the counts of the A-D cells at four times. In the short timeframe condition this happened between Trials 4-5, 12-13, 20-21, and after the causal strength judgment after Trial 24. In the long timeframe condition, this happened at the beginning of Days 5, 13, 21, and after the causal strength judgment on Day 25.

First, they recalled how many times the cause was present out of the trials that had been seen (e.g., "You have experienced [4, 12, 20, or 24] days. Out of these [4, 12, 20, or 24] days, how many days did you use Facebook?"). Suppose they answered 14. Their response was piped into two follow up questions to assess their memories for the number of times the effect was present when the cause was present (e.g., "Of the [14] days you did use Facebook, how many days were you in a very sad mood?") and how many days the effect was present when the cause was absent (e.g., "Of the [10] days you did not use Facebook, how many days were you in a very sad mood?").

From their responses, we were able to calculate their memories for cells of types A, B, C, and D of the contingency table.

2.3.3.3 Episodic Memories

Participants also made a number of judgments about the memories for the contextual images after Trial 24. These measures will not be analyzed in the current report.

2.3.4 Cover Stories

Since subjects learned about five cause-effect relations, we created five 'contexts' so that each task was viewed as a separate learning task (see Table 3). We chose these five cover stories such that it would be plausible for the cause to either improve or worsen the outcome. Because this study is the first to use a long timeframe paradigm, is unlikely to be replicated, and is focused on external validity, we manipulated two aspects of the cover stories: the "authenticity" and "valence". These manipulations were performed so that, we could rule out some potential explanations if subjects in the long timeframe condition exhibited very poor learning, and also to provide guidance for future studies with long timeframes.

For the matched short-term and long-term datasets, we assigned the same valence and authenticity conditions. Of the four short-term conditions, two had one version and two had the other version for both the valence and authenticity manipulations (see Table 2). Because these manipulations are not of primary importance, they are reported in the Appendix.

Table 3 Five Cover Stories used for Different Tasks

Authentic Cause and Context	Positive Valence	Negative Valence
Using Facebook during lunch in a restaurant	Very good mood vs. Normal mood	Very bad mood vs. Normal mood
Eating healthy dinner at a friend's house	Stomach feels great vs. Normal digestion	Upset stomach vs. Normal digestion
Using notecards to study for a quiz in a library	Very good grade vs. Normal grade	Very bad grade vs. Normal grade
Biking to work in city streets	Very productive vs. Normal productivity	Very unproductive vs. Normal productivity
Bringing your dog on a daily walk in the park	Very Relaxed vs. Feeling normal	Very Stressed vs. Feeling normal

Note. In the 'authentic' condition, participants learned about abstract relations. In the 'novel' condition, the contexts were the same, but the cause was replaced with a novel vitamin. The outcome for each cover story had either a positive or negative valence.

2.3.4.1 Authenticity vs. Novelty

We wanted to have cover stories that were both 'authentic' and 'novel'. Causal learning studies typically to use entirely novel cover stories (e.g., the effect of Medication X on sleep) to minimize the influence of prior beliefs, and we wanted to be able to replicate this typical paradigm. However, we also wanted to use authentic cover stories. We were worried that participants in the long timeframe condition might perform poorly with novel cover stories. In a short timeframe task, suspect subjects can often use other cues (e.g., the position of stimuli on the screen) rather than the semantic meanings of the cues. We worried that these alternative methods of learning might be less salient in the long timeframe condition. We thought that semantically meaningful cause-effect relations might be easier to remember and also have higher external validity.

The 'authentic' cover stories are listed in Table 3. In the 'novel' cover stories, we replaced the causes with a hypothetical vitamin that the subject took on some days but not others (e.g., does

Vitamin X have an influence on mood, upset stomach, etc.). For the matched short and long timeframe datasets, we matched the authenticity of the cause.

2.3.4.2 Valence

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 mirror 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. This is especially important because, valence can influence the strength of illusory correlations (Mullen & Johnson, 1990; Derringer, 2019). Thus, even though we are not primarily interested in valence, we counterbalanced the valence of the cover story.

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.

2.3.5 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.

In the long timeframe condition, the most important dependent measures 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.

3.0 Results

3.1 Measures of Strength

We analyzed three measures of strength: the causal strength question, the trial-by-trial predictions converted into a measure of strength, and the frequency judgments converted into a measure of strength. For all of these, we only analyzed data from the matched short-term and long-term conditions so that we could do within-subject tests. Tests were conducted separately for the generative (N = 98), preventive (N = 102), A-cell (N = 105), and outcome density (N = 104) conditions.

For each dataset and measure of causal strength, we conducted one-sample t-tests to see if short-term and long-term strength judgments were significantly different from zero. Next, we conducted paired-samples t-tests to assess whether there were significant differences between short-term and long-term judgments for each dataset. Because we observed non-normal distributions in the measures of strength, we also conducted non-parametric tests (Wilcoxon signed-rank). We also 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 BFs > 10 (or < 1/10) are considered "strong" evidence for the alternative (or null), BFs > 30 or < 1/30 are considered "very strong" and BFs >100 or < 1/100 are considered "extreme" (e.g., Lee & Wagenmakers, 2013).

3.2 Order Effects

Half of the participants completed the short-term task before, and half after the long-term task. To test for possible order effects in each measure of strength and timeframe condition, we conducted independent-sample *t*-tests between participants who saw the short-term task before the long-term task and participants who saw the short-term task after the long-term task. For example, we compared participants who did the short-generative task on Day 1 vs. participants who did the short-generative task on Day 25. We performed this t-test for each of the three measures of causal strength, and each of the four datasets, resulting in 12 tests. We also did a parallel analysis of the long-term judgments. For example, we compared the participants who did the long-generative task before the short-generative to those who did the long-generative after the short-generative, which produced another 12 tests.

Out of the 24 tests, only one was significant and only at p=.016. Because no systematic patterns in order effects appeared, and because we conducted 24 tests, so the likelihood of a Type I error is high, we concluded that there was no evidence for order effects. Thus, we did not change our strategy for analyzing strength judgments by doing within-subjects tests.

3.3 Causal Strength

Causal strength judgments were participants' explicit strength judgments after observing the 24 trials (e.g., "How strongly does Facebook improve your mood?"), transformed to be on a scale of -1 to +1 and reverse coded in the negative valence conditions. See Table 4 for frequentist, non-parametric, and Bayesian analyses.

3.3.1 Generative and Preventive Conditions

First, we wanted to assess whether participants were capable of detecting causation in the generative and preventive conditions (Figure 3A). Causal strength was significantly different from zero in both the short-term and long-term conditions for both the generative and preventive datasets, suggesting that participants did learn in both the short and long timeframe conditions.

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 or due to difficulty learning. However, the tests revealed no significant differences between judgments in the short-term and long-term conditions for either the generative or preventive datasets. In fact, the BFs were about 8 to 1 in favor of the null hypothesis. Thus, participants were just as capable of detecting causation in the short and long timeframe conditions.

3.3.2 Illusory Correlation Conditions

Consistent with our predictions, we found significant illusory correlations in both timeframe conditions for the A-cell and outcome-density datasets; both were greater than zero. We further hypothesized that the illusory correlations could be either exacerbated or diminished in the long timeframe condition. However, we did not find any differences between the short vs. long timeframe conditions. The BF for the A-cell condition was about 8 to 1 in favor of the null. There was a marginal effect for outcome-density but was not technically significant and the BF is actually slightly in favor of the null. Overall, these results suggest that illusory correlations in the long timeframe task are similar to the traditional trial-by-trial paradigm.

Table 4 Comparisons of Causal Strength Judgments for Short-Term and Long-Term Against Zero and for Short-Term vs. Long-Term

Dataset	*t*-test				Bayes Factor	Wilcoxon Test	
	t	df	p	d	BF	V	p
Short-Term							
Generative	11.27	97	<.001	1.14	$2.13*10^{16}$	3248	<.001
Preventative	-9.03	101	<.001	-0.89	$5.72*10^{11}$	193.0	<.001
A-cell	7.13	104	<.001	0.70	$6.75*10^{7}$	1273.5	<.001
Outcome Density	2.73	103	.008	0.27	3.60	317.0	.010
Long-Term							
Generative	11.53	97	<.001	1.17	$7.37*10^{16}$	2931.5	<.001
Preventative	-7.13	101	<.001	-0.71	$6.05*10^{7}$	252.0	<.001
A-cell	6.11	104	<.001	0.60	$6.36*10^{5}$	1684	<.001
Outcome Density	4.23	103	<.001	0.41	$3.41*10^{2}$	925.0	<.001
Short vs. Long							
Generative	-0.37	97	.707	-0.04	0.12	1358.0	.587
Preventative	-0.33	101	.741	-0.03	0.12	1605.5	.946
A-cell	-0.58	104	.563	-0.06	0.13	1386.0	.442
Outcome Density	-1.72	103	.089	-0.17	0.45	643.5	.068

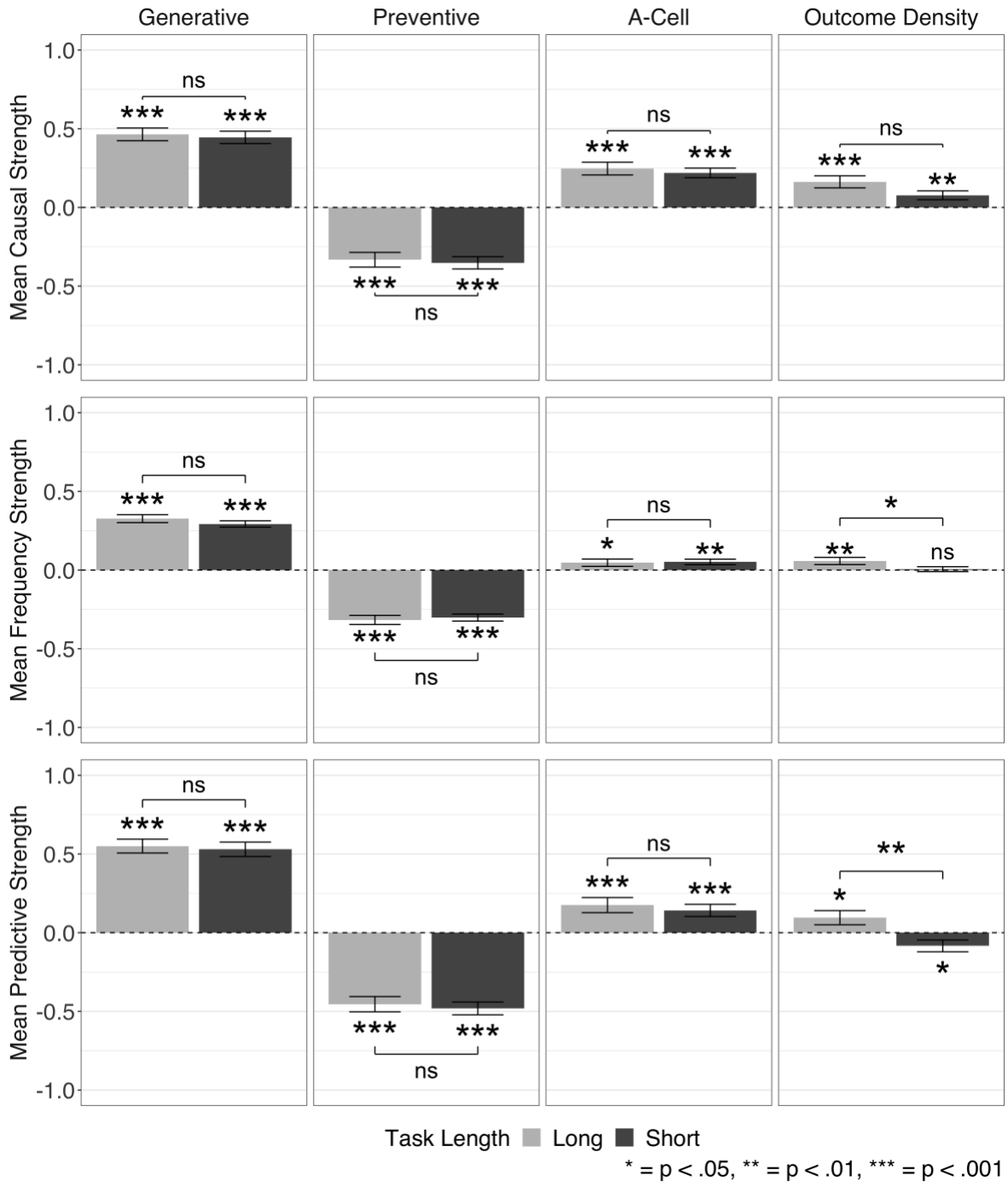

Figure 3 Average causal strength, predictive strength, and frequency strength judgments in the matched short-term and long-term conditions for the four datasets.

Significance markers above each column indicate whether the value was significantly different from zero. The significance marker above the horizontal lines indicates whether the judgments in the short and long-term conditions were significantly different from each other. Error bars indicate standard error.

3.4 Predictive Strength

We used participants trial-by-trial predictions of the effect to calculate a measure of their beliefs about the strength of the cause-effect relation which we call 'predictive strength'. We subtracted the probability that they predicted the outcome would be present given the absence of the cause from the probability that they predicted the outcome would be present given the presence of the cause. This measure is conceptually similar to ΔP: p(predict effect present|cause present) – p(predict effect present|cause absent). To ensure that participants had observed enough experiences to make predictions, we analyzed the predictions from Trials 13 – 24 (Table 5, Figure 3).

Table 5 Comparisons of Predictive Strength Judgments for Short-Term and Long-Term against Zero, and for Short-Term vs. Long-Term

Dataset	*t*-test				Bayes Factor	Wilcoxon Test	
	t	df	p	d	BF	V	p
Short-Term							
Generative	11.58	97	<.001	1.17	$9.01*10^{16}$	3254.5	<.001
Preventative	-11.87	101	<.001	-1.18	$6.77*10^{17}$	280.5	<.001
A-cell	3.66	104	<.001	0.36	51.08	3384.5	<.001
Outcome Density	-2.24	103	.028	-0.22	1.17	1476.0	.031
Long-Term							
Generative	12.47	97	<.001	1.26	$6.32*10^{18}$	4181.0	<.001
Preventative	-9.38	101	<.001	-0.93	$3.12*10^{12}$	405.5	<.001
A-cell	3.66	104	<.001	0.36	50.64	3448.5	<.001
Outcome Density	2.13	103	.036	0.21	0.94	2732.0	.036
Short vs. Long							
Generative	-0.36	97	.718	-0.04	0.12	1801.0	.634
Preventative	-0.49	101	.623	-0.05	0.12	2286.5	.747
A-cell	-0.66	104	.512	0.06	0.13	2278.5	.189
Outcome Density	-3.60	103	<.001	-0.35	42.20	1567.5	<.001

3.4.1 Generative and Preventive Conditions

We found very similar results using subjects' predictions to assess learning as from their causal strength judgments (see Figure 4). In the generative and preventive conditions, predictive strength was significantly different from zero for both the short-term and long-term conditions. Again, we found no difference in predictive strength between the short-term and long-term conditions for either the generative or preventive datasets.

3.4.2 Illusory Correlation Conditions

In the A-cell bias condition, we found a similar pattern of results to the causal 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 and long-term condition. Furthermore, we found no difference between predictions in the short-term vs. long-term conditions for the A-cell bias dataset.

We did observe a significant difference between short-term and long-term judgments for the outcome density dataset, although not in the direction we expected. In the long-term condition, we observed a small but significant illusory correlation. In the short-term condition, however, outcome-density judgments were negative and significantly different from zero in the short-term condition. Participants exhibited the opposite of an outcome density effect in the short-term condition in which predictive strength was negative and significantly different from zero. This pattern is inconsistent with prior rapid trial-by-trial studies and therefore, we are hesitant to interpret the significant difference between the short-term and long-term conditions.

3.5 Frequency Strength

We used participants memories of the frequencies of the four types of events to calculate a measure of their beliefs about the strength of the cause-effect relation which we call 'frequency strength' (Table 6, Figure 3). Frequency strength was calculated from subjects' memories for the frequencies of each cell type after Trial 24, using the ΔP formula: $A/(A+B) - C/(C+D)$.[1]

[1] In the outcome-density condition, we dropped one participant from analyses (see Table 6) because they gave judgments of $A = 0$, $B = 0$, $C = 0$, and $D = 24$, making it impossible to calculate frequency strength using ΔP and likely reflecting a desire to finish the task quickly.

Table 6 Comparisons of Frequency Strength Judgments for Short-Term and Long-Term against Zero, and for Short-Term vs. Long-Term

Dataset	*t*-test				Bayes Factor	Wilcoxon Test	
	t	df	p	d	BF	V	p
Short-Term							
Generative	14.46	97	<.001	1.46	$6.35*10^{22}$	4308.0	<.001
Preventative	-13.46	101	<.001	-1.33	$1.44*10^{21}$	140.0	<.001
A-cell	2.94	104	.004	0.29	6.13	3081.0	.001
Outcome Density	0.36	102	.718	0.04	0.12	2020.0	.521
Long-Term							
Generative	12.94	97	<.001	1.31	$5.83*10^{19}$	4321.0	<.001
Preventative	-10.92	101	<.001	-1.08	$6.27*10^{15}$	286.5	<.001
A-cell	3.66	104	<.001	0.36	0.74	2534.5	.030
Outcome Density	2.53	102	.013	0.25	2.23	2590.0	.016
Short vs. Long							
Generative	-1.21	97	.230	-0.12	0.23	1915.0	.232
Preventative	-0.46	101	.650	0.05	0.12	2639.0	.568
A-cell	-0.66	104	.512	0.06	0.11	2427.0	.997
Outcome Density	-3.60	103	<.001	-0.35	0.70	1871.0	.050

3.5.1 Generative and Preventive Conditions

Subjects' frequency strength followed the same pattern as we found in causal and predictive strength. Frequency strength was significantly different than zero for both timeframes and datasets, with no significant differences between the short-term and long-term conditions.

3.5.2 Illusory Correlation Conditions

For the A-cell dataset, we again found significant illusory correlations in both the short-term and long-term conditions with no significant difference between the two timeframes. For the outcome density dataset, frequency strength was significantly different from zero in the long-term

but not the short-term condition. There was no significant difference in frequency strength between the short-term and long-term conditions for the outcome-density. In sum, the results are extremely reliable except there for one condition – the long-term outcome-density condition. In this condition sometimes the judgments were greater than, sometimes less than, and sometimes roughly equal to zero. We briefly address potential explanations in the discussion section.

3.6 Measures of Strength over Time

In addition to the cumulative causal, predictive, and frequency strength judgments, we also collected data from these measures of strength at different points during the 24 learning trials (Figure 4). These judgments occurred every eight trials for the causal strength judgments (starting after Trial 8) and frequency strength judgments (starting after Trial 4). We calculated predictive strength every eight trials (starting on Trial 8), using the predictions participants made for the prior eight trials. We did not conduct quantitative analyses for the intermediary learning measures as we did for the cumulative ones to avoid problems associated with multiple comparisons given the number of tests for each measure of strength.

Qualitatively, the three measures of strength over time are quite similar between the short-term and long-term conditions. Figure 4 also shows that substantial learning has already occurred by Trial 8 for causal strength and predictive strength: the generative and preventive conditions are already separated. The frequency strength panel suggests that the long-term participants may have had a bit of difficulty remembering the prior evidence at Trial 4; the judgments are more extreme for the generative and preventative datasets in the short than long timeframe. However, by Trial 12 the short and long timeframes look similar.

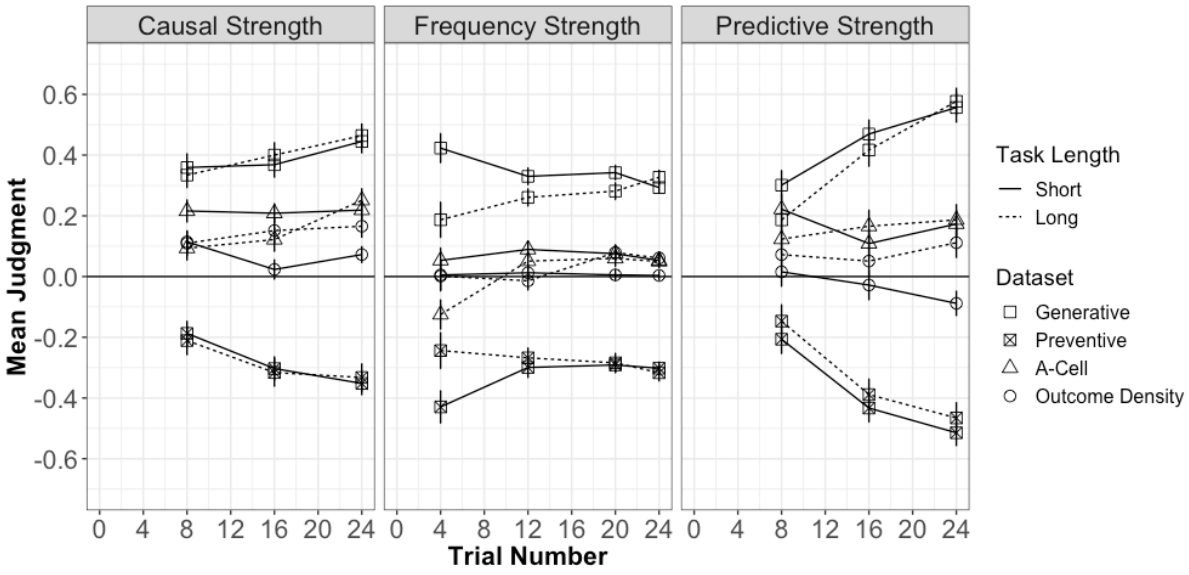

Figure 4 Measures of strength over time for each dataset in the short-term and long-term conditions.

4.0 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.

Overall, we found support for a null effect of task length on subjects' judgments across the three measures of causal strength. We did, however, observe some inconsistencies in judgments for the outcome density dataset. In the long-term conditions, causal, predictive, and frequency strength were positive and significantly different from zero, consistent with illusory correlations found in prior research. We also found a significant illusory correlation for causal strength judgments in the short-term condition. However, short-term predictive and frequency judgments for the outcome-density dataset were inconsistent – predictive strength was significantly negative (the opposite of an illusory correlation) and frequency strength was not significantly different from zero for the short-term task. In the predictive strength condition, negative judgments in the short-term version of the outcome-density dataset resulted in a small but significant difference between the short-term and long-term tasks.

Outcome density effects are pervasive in prior research, which exclusively used short-term designs (e.g., Allan et al., 2005; Buehner et al., 2003; Crump et al., 2007; Musca et al., 2010). Thus, we are unsure why we found some inconsistencies in the outcome-density condition. Our

primary goal for this study was to determine whether there are differences in the strength of illusory correlations in a long-term design. Given that the single significant effect of task length for the outcome density dataset is isolated to predictive strength and due to judgments that are contradictory with prior research, we suggest that there are little to no differences when using a short-term vs. long-term design. Furthermore, we found null effects for task-length differences across all measures of causal strength in the generative, preventive, and outcome-density datasets, supporting the external validity of short-term designs.

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. Using a large sample of participants, we found little to no differences in judgments between the short-term and long-term tasks. 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 all sorts of probability learning tasks 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? Though these theories are often extremely difficult to differentiate because they all predict similar learning, we can at least assess whether or not participants are able to recall episodic memories of the experienced events 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; Kello et al., 2010; Brown, Neath, & Chater; 2007), 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., Averell & Heathcote, 2010; Wixted & Ebbesen, 1991). These debates are complex and technical, and though the current study was not designed specifically to address either, perhaps researchers invested in those debates may be able to use incorporate 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.

Appendix A

Appendix A.1 Cover Story Authenticity

In both the matched short and long timeframe tasks, participants were randomly assigned to either a 'novel' (e.g., effect of taking Vitamin AQ46 on mood) or 'authentic' (e.g., effect of using Facebook on mood) cover story. We conducted separate ANOVA's using Type III error for each of the four datasets and three measures of causal strength to test for possible main effects of or interaction between cover story authenticity (between subjects) and task length (within subjects).

Out of the 12 significance tests of authenticity, only one was significant (Table A1). The causal strength judgments were stronger in the novel condition (M = 0.30, SD = 0.36) than in the authentic condition (M = 0.17, SD = 0.37) for the A-cell bias dataset. Thus, we conclude that there is no evidence of systematic effects of authenticity.

Out of the 12 significance tests of task length, only one test was significant; frequency strength illusory correlations were slightly stronger in the long-term task (M = 0.06, SD = 0.23) than in the short-term task (M = 0.00, SD = 0.16) for outcome density. This is the same finding already reported in the main paper.

Appendix Table 1 Effects of Coverstory Authenticity and Task Length on Measures of Strength

				Causal Strength		Predictive Strength		Frequency Strength	
Dataset	Predictor	df_{Num}	df_{Den}	F	p	F	p	F	p
Generative	Authenticity	1	190	1.28	.259	1.82	.178	0.02	.897
	Task Length	1	190	0.04	.849	1.00	.319	1.49	.224
	Interaction	1	190	2.66	.104	0.69	.407	1.71	.193
Preventive	Authenticity	1	198	0.53	.467	0.00	.967	1.30	.256
	Task Length	1	198	< 0.00	.950	0.45	.504	0.91	.341
	Interaction	1	198	0.01	.906	0.02	.884	0.00	.968
A-cell	Authenticity	1	206	6.38	.012*	1.01	.317	0.03	.871
	Task Length	1	206	1.38	.242	0.40	.528	0.20	.652
	Interaction	1	206	1.04	.309	0.39	.535	0.24	.623
Outcome Density	Authenticity	1	204	0.17	.684	0.27	.606	0.10	.751
	Task Length	1	204	0.24	.620	0.75	.387	3.95	.048*
	Interaction	1	204	2.15	.144	0.57	.451	1.83	.178

Note. * and boldface indicates that the effect was significant at $p < .05$.

Appendix A.2 Cover Story Valence

Prior research suggests that illusory correlations and strength judgments may be stronger for outcomes that have negative than positive valence (Derringer, 2019; Mullen & Johnson, 1990). There is also some evidence that people may learn generative and preventive relations faster or give stronger judgments for negative than positive valenced outcomes (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Ohman & Mineka, 2001; Rozin & Royzman, 2001).

In both the matched short and long timeframe tasks, the present effect had either a positive or negative valence (e.g., ‘Very Good Mood’ vs. ‘Very Bad Mood’). We conducted separate ANOVA’s using Type III error for each of the four datasets and three measures of causal strength to test for possible main effects of or interaction between cover story valence (between subjects) and task length (within subjects) (Table A2).

Consistent with our authenticity analyses, we found a single, barely-significant effect of task length that was isolated to frequency strength for the outcome-density bias dataset.

For the illusory correlation datasets, only one of the six tests was significant and was in the opposite direction as predicted; the predictive strength measure for outcome density was slightly higher for the positive ($M_{positive}$ = 0.06, $SD_{positive}$ = 0.45) than negative ($M_{negative}$ = -0.05, $SD_{negative}$ = 0.41) condition.

For the causal datasets, the findings were mixed. For the preventive dataset, we observed a pattern consistent with prior research in which predictive strength was significantly stronger in the negative valence condition (M = 0.63, SD = 0.38) than in the positive valence condition (M = 0.45, SD = 0.49). However, for the generative dataset, predictive strength was significantly stronger for the positive than the negative valence condition in both the generative ($M_{positive}$ = 0.63, $SD_{positive}$ = 0.38, $M_{negative}$ = 0.45, $SD_{negative}$ = 0.49). Three of other findings for the other measures of strength were non-significant, and one was marginal.

In sum, unlike some prior studies, we did not see strong and reliable patterns effects of valence for either the short or long timeframes.

Appendix Table 2 Effects of Coverstory Valence and Task Length on Measures of Strength

				Causal Strength		Predictive Strength		Frequency Strength	
Dataset	Predictor	df_{Num}	df_{Den}	F	p	F	p	F	p
Generative	Valence	1	190	0.29	.589	7.84	.006*	1.39	.240
	Task Length	1	190	0.04	.850	1.04	.310	1.49	.223
	Interaction	1	190	< 0.00	.979	2.28	.133	0.73	.394
Preventive	Valence	1	198	3.75	.054	5.63	.019*	1.30	.256
	Task Length	1	198	< 0.00	.951	0.46	.498	0.91	.341
	Interaction	1	198	0.05	.830	0.17	.683	< 0.00	.964
A-cell	Valence	1	206	0.14	.705	2.03	.156	1.98	.161
	Task Length	1	206	1.33	.240	0.39	.535	0.21	.650
	Interaction	1	206	0.40	.526	0.90	.343	1.28	.260
Outcome	Valence	1	204	0.49	.486	4.58	.034*	0.53	.468
Density	Task Length	1	204	0.25	.620	0.77	.383	3.95	.048*
	Interaction	1	204	1.41	.237	0.03	.857	1.53	.217

Bibliography

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15*(3), 147-149.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology/Revue anadienne de psychologie*, *34*(1), 1-11.

Allan, L. G., Siegel, S., & Tangen, J. M. (2005). A signal detection analysis of contingency data. *Learning & Behavior, 33*(2), 250-263.

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology, 55*(1), 25-35.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370.

Blanco, F., Matute, H., & Vadillo, M. A. (2013). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & Behavior, 41*(4), 333-340.

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*(7), 997-1003.

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539-576.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1119-1140.

Carranza-Jasso, R., Urcelay, G. P., Nieto, J., & Sánchez-Carrasco, L. (2014). Intertrial intervals and contextual conditioning in appetitive Pavlovian learning: Effects over the ABA renewal paradigm. *Behavioural Processes, 107*, 47-60.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236-246.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354-380.

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.

Crump, M. C., Hannah, S. D., Allan, L. G., & Hord, L. K. (2007). Contingency judgments on the fly. *The Quarterly Journal of Experimental Psychology, 60*(6), 753-761. doi:10.1080/17470210701257685

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A., Robbins, T. W., Gasull-Camos, J., ... & Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General*, *147*(7), 1043-1065.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876-879.

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, *84*(6), 3072-3077.

Derringer, C. (2019). Illusory Correlation and Valenced Outcomes. Unpublished doctoral dissertation. University of Pittsburgh, Pittsburgh, PA.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795-805.

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

Grogan, J. P., Tsivos, D., Smith, L., Knight, B. E., Bogacz, R., Whone, A., & Coulthard, E. J. (2017). Effects of dopamine on reinforcement learning and consolidation in Parkinson's disease. *Elife, 6*, e26801.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334-384.

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.

Holland, P. C., & Morell, J. R. (1996). The effects of intertrial and feature–target intervals on operant serial feature negative discrimination learning. *Learning and Motivation*, *27*(1), 21-42.

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.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences, 14*(5), 223-232.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22-44.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron, 20*(5), 937-945.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32*(2), 311-328.

Le Pelley, M. E., Reimers, S. J., Calvini, G., Spears, R., Beesley, T., & Murphy, R. A. (2010). Stereotype formation: Biased by association. *Journal of Experimental Psychology: General, 139*(1), 138-161.

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

Metcalfe, J., & Xu, J. (2016). People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(6), 978-984.

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

Musca, S. C., Vadillo, M. A., Blanco, F., & Matute, H. (2010). The role of cue information in the outcome-density effect: Evidence from neural network simulations and a causal learning experiment. *Connection Science, 22*(2), 177-192.

Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. E. (2005). Depressive realism and outcome density bias in contingency judgments: the effect of the context and intertrial interval. *Journal of Experimental Psychology: General, 134*(1), 10-22.

Mustaca, A. E., Gabelli, F., Papini, M. R., & Balsam, P. (1991). The effects of varying the interreinforcement interval on appetitive contextual conditioning. *Animal Learning & Behavior, 19*(2), 125-138.

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.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483-522.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory, 2*, 64-99.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296-320.

Schiller, D., Monfils, M. H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature, 463*(7277), 49-53.

Sheffeld, V. F. (1950). Resistance to extinction as a function of the distribution of extinction trials. *Journal of Experimental Psychology*, *40*(3), 305-313.

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

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.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron, 90*(5), 927-939.

Stanley, W. C. (1952). Extinction as a function of the spacing of extinction trials. *Journal of Experimental Psychology, 43*(4), 249-260.

Tricomi E, Balleine BW, O'Doherty JP (2009) A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience, 29,* 2225– 2232.

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psychology, 53*(1), 1-25.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109*, 163-167.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8*(3), 600-608.

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

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. *Journal of Neuroscience, 38*(35), 7649-7666.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience, 15*(5), 786-791.