The Accuracy of Causal Learning over Long Timeframes: An Ecological Momentary Experiment Approach

Ciara L. Willett
Benjamin M. Rottman

1 Introduction

The ability to learn relations between events from experience is critical for humans to behave adaptively. Learning relations between events allows individuals to predict future events (e.g., predicting the likelihood of a side-effect after taking a medicine) and make decisions to try to bring about desirable outcomes (e.g., deciding whether or not to use the medication). There are two types of errors when learning relations between events. First, one might falsely conclude that there is no relation between two events. For example, if a person stops taking a medication because they falsely judge it to be ineffective, they will miss out on the benefits of the therapy. Another error, called *illusory correlation* or *illusory causation*, involves falsely concluding that there is a relation between two events. This type of error is believed to contribute to the formation of stereotypes (e.g., believing that one race is more likely to commit crimes than another (Hamilton & Gifford, 1976) and believing in pseudoscientific therapies (Matute et al., 2011)). The goal of the present research was to assess whether and when people can accurately identify relations, which is crucial for informing efforts to improve decision making.

These questions have frequently been studied using the "trial-by-trial" paradigm in which participants observe multiple cue-outcome (or 'cause-effect') pairs. This trial-by-trial paradigm simulates how individuals learn relations through trial and error while experiencing a temporal stream of events. Typically, the entire learning session lasts on the order of 5-10 minutes (or 30-60 minutes for tasks with many trials). Furthermore, the participant is fully engaged in the learning process and any distractions during the task are experimenter-induced. Originally used in the behaviorist tradition, trial-by-trial learning is used pervasively across many fields such as causal learning (Spellman, 1996; Waldmann, 2001), correlation detection (Jenkins & Ward, 1965; Kao & Wasserman, 1993), reinforcement learning (Daw et al., 2006; Delgado et al., 2000), category learning (Kruschke, 1992; Nosofsky, 1986), fear learning (LaBar et al., 1998; Schiller et al., 2010), and stereotype formation (Hamilton & Gifford, 1976; Le Pelley et al., 2010).

However, we contend that this pervasive rapid trial-by-trial learning paradigm does not reflect real-world learning situations, which usually involve experiencing cause-effect associations over longer periods of time. For example, stereotypes are not learned on the order of minutes, but through experiences with in-group and out-group members over longer periods of time. Similarly, learning about potential food allergies or the effectiveness of a medication is based on experiences that are spaced out over days or weeks. When learning is spread out over time and embedded in daily life, the learner is simultaneously engaged in many other cognitive processes. Because the learner must rely on long-term memory as opposed to working memory, they may have more difficulty learning cause-effect relations and be more likely to make mistakes in real-world situations.

Historically, because so much research on causal learning (and experience-based learning in general) compares human learning and judgments to normative or rational computational models, researchers have often studied learning under optimal situations in which memory demands are minimized. In the trial-by-trial paradigm, this is accomplished by rapidly presenting the trials back-to-back usually without distractions, and in some other paradigms the data is presented in a summarized format to make it even easier to digest (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005). Because of this emphasis on studying learning and reasoning under optimal conditions, fairly little research has investigated learning under more real-world learning constraints, though there are a few exceptions.

In fact, two of the foundational studies on illusory correlations investigated learning that naturally occurred over a spaced-out timeframe. Chapman and Chapman (1969) argued that psychotherapists inferred illusory correlations between Wheeler-Rorschach signs and their patients' diagnoses, and Redelmeier and Tversky (1996) argued that arthritis patients experience illusory correlations between the weather and their symptoms. However, both of these assessments are flawed. Aside from the premise of diagnosing homosexuality being highly problematic (Herek, 2010), Chapman and Chapman also did not have access to the individual therapists' clinical records, so they could not actually analyze learning. And Redelmeier and Tversky did not take autocorrelation into account when analyzing correlations

between weather and pain, which is necessary for analyzing time-series data. In sum, there is no prior research that can really speak to the accuracy of causal learning over long timeframes.

1.2 The Role of Memory in Causal Learning

We compared trial-by-trial learning in the standard rapid paradigm with 24 trials (short timeframe) versus learning in which the trials were spaced out once per day for 24 days (long timeframe). The long timeframe condition, in which one trial was experienced per day, was intended to simulate natural processes that unfold on a daily timescale (e.g., does a medicine that can be taken once per day influence a health outcome, does exercising on some days influence sleep).

We investigated how well people learned about four datasets. In two datasets, there was a real correlation between the cause and the effect, either positive or negative. In two datasets, there was zero correlation between the cause and the effect; however, prior research has shown that people typically infer illusory correlations for these.

None of the existing models of causal learning have memory built into the model, so they do not make clear predictions about learning over spaced-out timeframes. We briefly discuss these models to point towards potential predictions that they could make. Importantly, our goal is not to confirm or disconfirm particular models; that is impossible because they do not make clear predictions. Still, it is possible to hypothesize a range of predictions that these models could make for long timeframes.

Rule-based theories of causal learning (Cheng, 1997; Griffiths & Tenenbaum, 2005; Hattori & Oaksford, 2007) assume that people judge causal relationships from tallies of the experienced events. If learning over a long timeframe causes increased interference and/or decay, and if this leads to less accurate tallies of the experienced events, it would mean worse judgments in the long timeframe compared to the short. For example, in studies that have manipulated working memory load, illusory correlations become stronger and the ability to detect true relations is impaired when working memory is taxed (Kao & Wasserman, 1993; Shaklee & Mims, 1982). Additionally, older adults with working memory decline and people with lower levels of working memory exhibit stronger illusory correlations compared to younger adults and people with higher levels of working memory (Eder et al., 2011; Mutter &

Pliske, 1996). In sum, perhaps the current study, which stretches learning out over time and therefore requires long term memory instead of short term memory, could be viewed somewhat analogously to studies that have increased working memory load.

If memory is worse in the long timeframe, there are a few possible predictions. For the datasets with a true cause-effect correlation, worse learning would likely mean that learners would have more difficulty detecting the statistical relationship in the long timeframe compared to the short, so judgments would be closer to zero. For the illusory correlation datasets, worse learning could play out in two ways. First, it is possible that there will be stronger illusory correlations in the long timeframe condition similar to the studies that manipulated working memory load (Kao & Wasserman, 1993; Shaklee & Mims, 1982). Second, if people have a lot of difficulty learning the statistical relationship, it could lead to judgments closer to zero, which would paradoxically produce more accurate judgments.

Alternatively, there may be few differences between the long vs. short timeframe conditions. In regard to the rule-based theories, it is possible that people will have fairly accurate memories of the tallies given that they only involve a fairly simple form of learning between a single cause and a single effect. Associative and reinforcement-learning models (e.g., Rescorla & Wagner, 1972) might also predict fairly accurate judgments, given that they do not require memories for individual events and only require sequentially updating a cue-outcome association.

Because the existing models do not have memory built in, they do not make clear predictions for causal learning over long timeframes. Still, it is vitally important to know whether people can accurately assess statistical relations in their own lives such as these daily processes. As the first study of its kind, this study was designed to provide strong evidence about the accuracy of simple causal learning over longer timeframes.

2 Materials and Methods

2.1 Participants

There were 479 participants (mean age = 21 years, 96% under 30 years old). Participants were required to own a smartphone and intend to complete the entire study. 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. This study was approved as exempt through our university's IRB. All participants gave informed consent. Our goal was to have around 400 participants, 100 for each of the 4 datasets in the long timeframe condition. The final data analyses included 413 participants after dropping 13 who admitted to writing down data during the study, 1 who admitted to not trying during the task, 40 due to a programming error, and 12 who skipped too many days.

2.2 Datasets and Design

Participants learned the statistical relation between a binary cause (*c*) and a binary effect (*e*) in four datasets, each with 24 randomly ordered observations (Table 1). We represent the presence vs. absence of the cause and effect with 1 and 0 respectively. We used the Δ*P* rule (Allan, 1980) to characterize the strength of the cause-effect relations in the datasets, Δ*P* = *P*(*e*=1|*c*=1) – *P*(*e*=1|*c*=0), as well as Cheng's (1997) power PC metric of causal strength (Table 1).

Table 1. Frequencies for Datasets in Current Study

Dataset	*c*=1 *e*=1	*c*=1 *e*=0	*c*=0 *e*=1	*c*=0 *e*=0	*P*(*e*=1|*c*=1)	*P*(*e*=1|*c*=0)	Δ*P*	Power PC
Generative	9	3	3	9	.75	.25	.5	.67
Preventive	3	9	9	3	.25	.75	-.5	-.67
A-cell	10	6	5	3	.625	.625	.0	.00
Outcome-Density	9	3	9	3	.75	.75	.0	.00

Note. *C* represents the cause and *e* represents the effect.

In the "generative" dataset, there was a positive relation between the cause and effect, and in the "preventive" dataset there was a negative relation. Both Δ*P* and power PC imply moderate strength relations for the generative and preventive datasets. People typically make roughly normative judgments for these datasets in short timeframe studies (e.g., Shaklee & Mims, 1982).

There were two noncontingent datasets. The A-cell dataset had a high number of 'A-cell' (*c*=1, *e*=1) trials, and the 'outcome density' dataset had a high probability of the effect. Both Δ*P* and power PC imply zero relations between the cause and outcome for these two datasets,

however, people typically infer a positive relation for these datasets, which is called an "illusory correlation" (e.g., Blanco et al., 2013; Kao & Wasserman, 1993).

Participants completed five tasks (Table 2). They learned about one of the four datasets in the long timeframe, and they learned about all four in the short timeframe. On Day 1 they completed Tasks 1 and 2 – the short timeframe tasks for 2 of the 4 datasets. Then from Days 1-24 they completed Task 3 - the long timeframe condition. Finally, on Day 25, they completed Tasks 4 and 5 – the short timeframe tasks for the 2 datasets not experienced on Day 1.

The long timeframe dataset matched one of the four short timeframe datasets in all ways except for the length and the context of the cover story. Having subjects learn all four datasets in the short timeframe condition was done to permit a within-subjects comparison for increased power for this hard to run and expensive study while reducing the likelihood that subjects were aware that one of the short timeframe datasets was the same as the long timeframe dataset. The short timeframe dataset that matched the long timeframe dataset was randomly assigned to appear either on Day 1 or Day 25. Of the two short timeframe tasks on Days 1 and 25, one was contingent (generative or preventive) and one was noncontingent (outcome density or A-cell). Table 2 gives an example of the design for one participant.

Table 2. Example order of tasks and randomization for a single subject

Task Order	Day	Length	Dataset	Context	Valence	Authenticity
1	1	Short	A-cell*	Restaurant	Positive*	Authentic*
2	1	Short	Preventive	House	Negative	Novel
3	1-24	Long	A-cell*	Library	Positive*	Authentic*
4	25	Short	Generative	Street	Positive	Novel
5	25	Short	Outcome density	Park	Negative	Authentic

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.

2.2.1 Cover stories. Since subjects learned about five cause-effect relations, we created five 'contexts' and cover stories so that each task was viewed as a separate learning task. In each cover story, it was plausible for the cause to either improve or worsen the effect.

Out of caution with this new and resource-intensive paradigm, we manipulated two aspects of the cover stories: authenticity of the cause (authentic vs. novel) and valence of the effect (positive vs. negative). Typical causal learning paradigms use novel causes (e.g., the effect of Vitamin E8H9 on productivity) to minimize the influence of prior beliefs. However, our overall goal for this study was to increase the external validity (primarily in terms of timeframe, but also in terms of feeling real more generally) and were worried that participants in the long timeframe condition might perform poorly with novel causes (Mutter & Plumlee, 2009). Thus, we decided to use both authentic and novel cover stories (see Table 3 for the authentic cover stories). In the novel cause condition, the causes involved taking a hypothetical vitamin or not.

Most effects have an implicit valence of being good (e.g., flower blooming; Spellman, 1996) or bad (e.g., presence/absence of a headace; Liljeholm & Cheng, 2007). Furthermore, valence can lead to stronger judgments about illusory correlations (Bott & Meiser, 2020; Mullen & Johnson, 1990) and real correlations (Baumeister et al., 2001; Öhman & Mineka, 2001; Rozin & Royzman, 2001). For generality, we used both negative and positive valence. The absence of the effect was always described as “normal”. The presence of the effect was described as either very good or very bad (Table 3). We coded the causal strength judgments so positive causal strength always means a positive correlation between the presence of the cause and the presence of the effect.

We assigned the same valence and authenticity conditions to the matched short timeframe and long timeframe datasets. Of the four short timeframe conditions, two had one version and two had the other version for both the valence and authenticity manipulations (Table 2).

The valence and authenticity manipulations were intended mainly for exploratory and counterbalancing purposes. We found no systematic effects of authenticity or valence and report these results in the supplement.

Table 3. Authentic causes, contexts, and effect valences for cover stories

Authentic Cause	Contextual Image	Effect: Positive Valence	Effect: Negative Valence
Using Facebook vs. Not using Facebook	Restaurant	Very good mood vs. Normal mood	Very bad mood vs. Normal mood
Eating a healthy dinner vs. Not eating a healthy dinner	Friend's house	Stomach feels great vs. Normal digestion	Upset stomach vs. Normal digestion
Using notecards to study vs. Not using notecards to study	Library	Very good grade vs. Normal grade	Very bad grade vs. Normal grade
Biking to work vs. Not biking to work	Streets	Very productive vs. Normal productivity	Very unproductive vs. Normal productivity
Bring dog vs. Not bringing dog on a daily walk	Park	Very relaxed vs. Feeling normal	Very stressed vs. Feeling normal

Note. In the 'novel' condition, the cause was replaced with taking a novel vitamin (e.g., Vitamin E8H9) or not.

2.3 Procedure

2.3.1 Overall procedure. The study was run on a website created with our PsychCloud.org framework (Rottman, 2019) and participants used their personal smartphones. The procedure for the short timeframe and long timeframe tasks were identical, except that there was one trial per day in the long timeframe condition and the trials were back-to-back in the short timeframe condition. On Day 1 of the study, participants completed two short timeframe tasks and began Day 1 of the long timeframe task.

The long timeframe task occurred on Days 1-24. Participants received automated text-message reminders at 10am, 3pm, and 8pm to complete the trial. They stopped receiving reminders if they had already participated that day. Participants were told that if they missed more than three days, the study would be terminated and that they would not be paid; 464 (97%) completed the study. On any given day, 81% participated before the 3pm reminder, 95% before the 8pm reminder, and 98% by midnight. For subjects who failed to participate on one (13%), two (5%), or three days (2%), the subsequent trials were automatically pushed back the appropriate number of days so that they experienced all 24 trials.

They returned to the lab on Day 25 to answer questions about the long timeframe task, complete the remaining two short timeframe tasks, and receive payment. Of the 413 subjects in the final analyses, 83% returned to the lab the day after completing Trial 24. Sometimes participants returned to the lab on the same day as Trial 24 (13%), or two (2%), three (1%), or four (.2%) days after Trial 24, depending on the number of missed days and their availability.

In sum, the long-timeframe study protocol of one trial per day for 24 days, with assessments on Day 25, was followed with high fidelity.

2.3.2 Within a trial. Each trial proceeded similarly to the following example (Fig. 1). In this example, subjects were asked to judge whether taking Vitamin E8H9 during their lunch break improves or worsens or has no influence on their mood, based on the hypothetical data presented to them (not participants' real-life experiences).

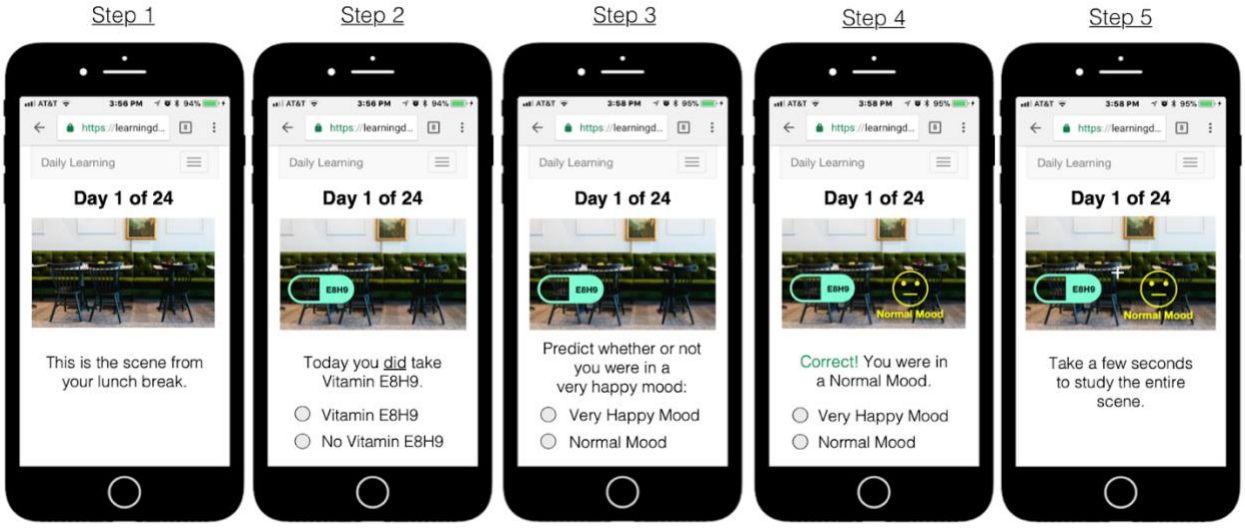

Figure 1. Screenshots depicting one trial. Text has been simplified for visibility.

In Step 1, subjects were shown a contextual image and description for three seconds. In Step 2, an icon and text appeared to show the presence or absence of the cause and participants confirmed the state of the cause. In Step 3, they predicted the effect as present or absent. In Step 4, they received feedback, saw an icon representing the true state of the effect, and verified the state of the effect. In Step 5, they were instructed to study the scene for four seconds.

At the end of a trial in the short timeframe condition, subjects moved 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. In both the short and long timeframe conditions, participants were unable to access a trial after completing it; they could not go back to see what happened on a previous trial/day at any point during the study.

2.4 Dependent Variables

2.4.1 Causal strength. Before Trials 9, 17, and after Trial 24, participants answered whether the cause (Vitamin E8H9) "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" (+1) or "worsen" (-1), they answered "How strongly does [the cause] [improve/worsen] [the effect]?" on a scale of 1 (very weak) to 10 (very strong). These two questions were multiplied together and divided by 10 to produce a causal strength rating from -1 (negative relation between case and effect) to +1 (positive relation between cause and effect).

2.4.2 Frequency judgments. Before Trials 5, 13, 21, and after the causal strength judgment on Trial 24, participants recalled how often they experienced four types of events (each combination of cause and effect, as present or absent) using the following three questions. First, they recalled how many times the cause was present out of the trials that had been seen (e.g., "You have experienced 24 days. Out of these 24 days, how many days did you take Vitamin E8H9?"). The participant's response was piped into two follow-up questions: "Of the [14] days you did take Vitamin E8H9, how many days were you in a very happy mood?" and "Of the [10] days you did not take Vitamin E8H9, how many days were you in a very happy mood?".

2.4.3 Memory task. After the frequency judgments, participants completed a three-part memory task. First, in the *recognition memory task*, they were asked to choose which of two images they saw during the learning task (one image they had previously seen and a lure image). Participants were then given feedback about the actual image they saw during the learning task. Second, in the *episodic memory task*, participants recalled whether the cause and effect were present or absent for the image that they saw during the task. Finally, in the *temporal order task*, participants used a slider to report the approximate trial (1-24) during which the scene occurred. Only the results from the episodic memory task are reported in the manuscript.

3 Results

The data and analysis scripts are available at https://osf.io/hmzvn.

3.1 Analysis Plan

We analyzed three measures of participants' judgments for the strength of the relation between the cause and effect at the end of learning: the causal strength question, the trial-by-trial predictions, and the frequency judgments (Fig. 2). We converted the predictions and frequency judgments into measures of the strength of the cause-effect relation using the ΔP equation[1]. For predictive strength, we used the predictions from Trials 13-24 to ensure that participants had some time to learn the cause-effect relation.

[1] For frequency ratings in the outcome-density condition, we dropped one participant who said that there were 24 *c*=0, *e*=0 trials, making *Δ*P incalculable.

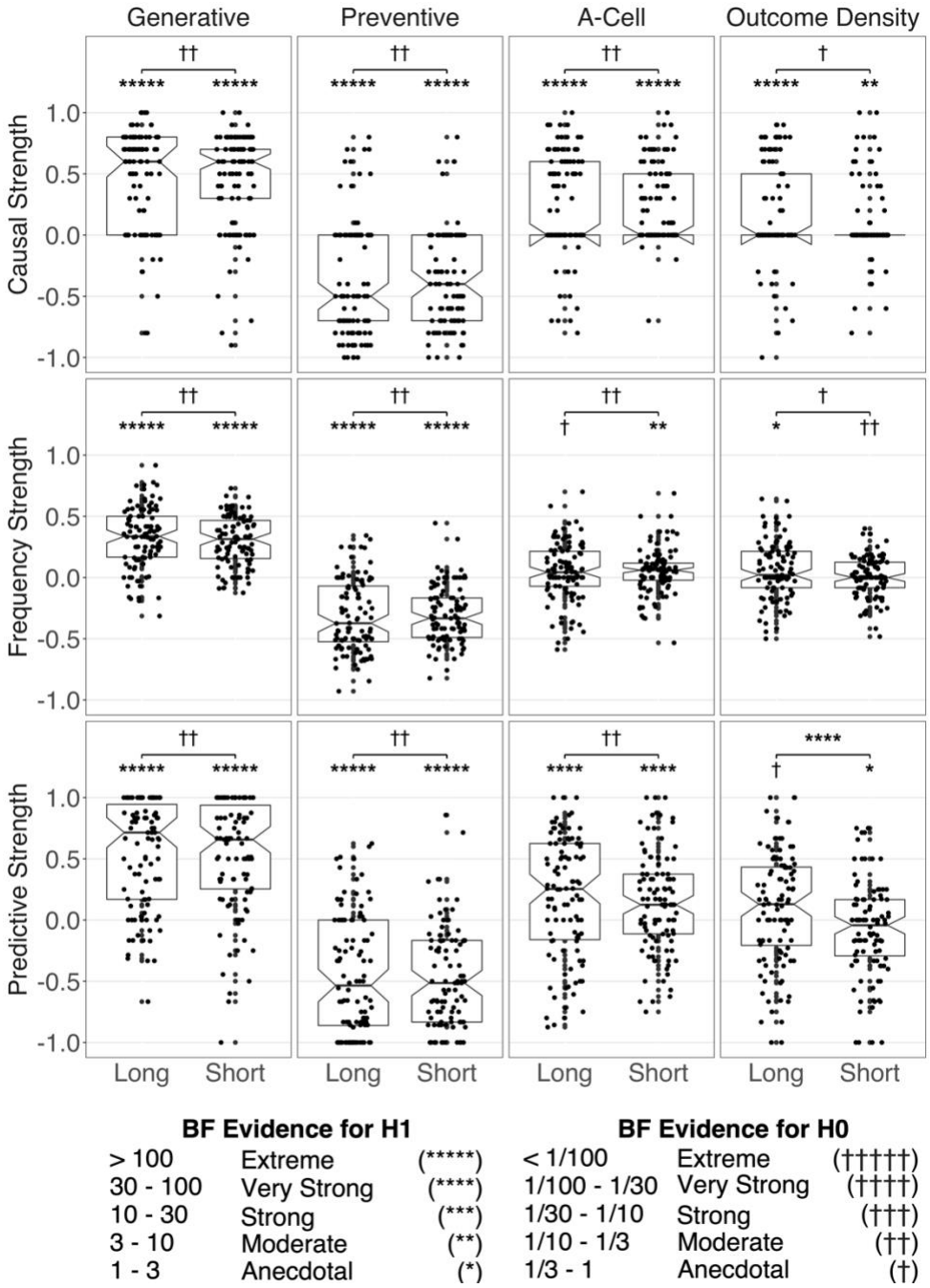

Figure 2. The box plot shows the 25th, median, and 75th percentiles, and notches represent the 95% CI of the median. Individual observations (*N* = 413) were plotted with horizontal jitter. Markers above each column indicate Bayes Factor evidence for the value vs. zero. Markers above the horizontal lines indicate Bayes Factor evidence for differences between the short vs. long timeframe conditions.

We only analyzed data from the matched short timeframe and long timeframe conditions so that we could do within-subject tests. Tests were conducted separately for the generative (*N* = 99), preventive (*N* = 104), A-cell (*N* = 106), and outcome density (*N* = 104) conditions.

Participants were randomly assigned to complete the matched short condition before (on Day 1) or after the long condition (on Day 25). To analyze for possible order effects, in the short timeframe condition, we compared participants who did the short timeframe task on Day 1 vs. participants who did the short timeframe task on Day 25. We did the same analysis for the long timeframe condition. This resulted in 24 t-tests; one for each of the 2 timeframes, 3 measures, and 4 datasets. There was only one significant result (see Appendix for all results). In the short timeframe, predictive strength judgments for the preventive dataset were significantly stronger if learned after (*M* = -0.57, *SD* = 0.37) compared to before (*M* = -0.38, *SD* = 0.43) the long task, $t(102) = 2.37$, $p = .020$, $d = 0.46$. Because only one of 24 t-tests was significant, we concluded that there was no reliable evidence of systematic order effects and followed our intended analysis plan.

The inferential statistics for the causal strength, prediction strength, and frequency strength measures appear in Tables 4 and 5. For each dataset and measure of causal strength, we conducted two-sided one-sample t-tests against zero to see if short timeframe and long timeframe strength judgments were significantly different from zero. Next, we conducted two-sided paired-samples t-tests to assess whether there were significant differences between short timeframe and long timeframe judgments for each dataset. We also calculated Bayes Factors (BF) for each t-test using the default parameters and priors in the BayesFactor package (Morey & Rouder, 2018) in R including the default sqrt(2)/2 prior for the rscale parameter. A BF > 1 is support for the alternative hypothesis and a BF < 1 is support for the null. We used the descriptive labels to summarize the BFs suggested by Lee and Wagenmakers (2014). Because we observed non-normal distributions in the measures of strength, we also conducted non-parametric Wilcoxon signed-rank tests. The conclusions were very similar across the t-tests, BFs, and Wilcoxon tests.

Table 4. Results for Generative and Preventive Conditions

	t-test					Bayes Factor	Wilcoxon Test	
	t	*p*	*d*	*CI lower*	*CI upper*	*BF*	*V*	*p*
Generative (N=99)								
Long vs. 0								
Causal	11.53	<.001	1.18	.39	.55	1.84×10^{17}	3.01×10^{3}	<.001
Frequency	12.97	<.001	1.30	.28	.37	7.79×10^{19}	4.41×10^{3}	<.001
Predictive	12.37	<.001	1.24	.46	.63	4.67×10^{18}	4.18×10^{3}	<.001
Short vs. 0								
Causal	11.43	<.001	1.15	.37	.53	5.25×10^{16}	3.33×10^{3}	<.001
Frequency	14.61	<.001	1.47	.26	.34	1.58×10^{23}	4.40×10^{3}	<.001
Predictive	11.65	<.001	1.17	.44	.62	1.51×10^{17}	4.36×10^{3}	<.001
Long vs. Short								
Causal	-0.38	.707	-0.05	-.12	.08	0.12	1.36×10^{3}	.585
Frequency	-1.04	.301	-0.13	-.09	.03	0.19	2.00×10^{3}	.299
Predictive	-0.29	.769	-0.04	-.13	.09	0.12	1.87×10^{3}	.705
Preventive (N=104)								
Long vs. 0								
Causal	-7.24	<.001	-0.71	-.42	-.24	1.10×10^{8}	2.70×10^{2}	<.001
Frequency	-11.04	<.001	-1.08	-.37	-.26	1.47×10^{16}	2.99×10^{2}	<.001
Predictive	-9.44	<.001	-0.93	-.55	-.36	5.04×10^{12}	4.27×10^{2}	<.001
Short vs. 0								
Causal	-9.21	<.001	-0.90	-.43	-.28	1.57×10^{12}	2.03×10^{2}	<.001
Frequency	-13.75	<.001	-1.35	-.35	-.26	8.26×10^{21}	1.41×10^{2}	<.001
Predictive	-12.15	<.001	-1.19	-.56	-.41	3.58×10^{18}	2.84×10^{2}	<.001
Long vs. Short								
Causal	-0.35	.728	-0.05	-.14	.09	0.12	1.69×10^{3}	.954
Frequency	0.25	.801	0.03	-.06	.07	0.11	2.68×10^{3}	.718
Predictive	-0.62	.538	-0.07	-.14	.07	0.13	2.34×10^{3}	.628

Note. Long vs. 0 and Short vs. 0 are statistical tests to see if the mean or median is different from zero. The Long vs. Short comparison subtracts long from short, so positive numbers mean that short is more positive and negative means that short is more negative.

Table 5. Results for A-Cell and Outcome Density Conditions

	t-test					Bayes Factor	Wilcoxon Test	
	t	*p*	*d*	*CI lower*	*CI upper*	*BF*	*V*	*p*
A-Cell (N = 106)								
Long vs. 0								
Causal	6.24	<.001	0.61	.17	.33	1.15×10^{6}	1.74×10^{3}	<.001
Frequency	1.96	.052	0.19	.00	.09	0.68	2.57×10^{3}	.035
Predictive	3.79	<.001	0.37	.09	.28	7.56×10^{1}	3.54×10^{3}	<.001
Short vs. 0								
Causal	7.12	<.001	0.69	.16	.28	6.46×10^{7}	1.27×10^{3}	<.001
Frequency	2.88	.005	0.28	.02	.09	5.31	3.12×10^{3}	.002
Predictive	3.75	<.001	0.36	.07	.22	6.74×10^{1}	3.47×10^{3}	<.001
Long vs. Short								
Causal	-0.71	.477	-0.09	-.13	.06	0.14	1.39×10^{3}	.364
Frequency	0.20	.845	0.02	-.05	.06	0.11	2.48×10^{3}	.985
Predictive	-0.72	.469	-0.08	-.14	.06	0.14	2.30×10^{3}	.161
Outcome Density (N=104, N=103 for Frequency Strength)								
Long vs. 0								
Causal	4.23	<.001	0.41	.09	.24	3.41×10^{2}	9.25×10^{2}	<.001
Frequency	2.53	.013	0.25	.01	.10	2.23	2.59×10^{3}	.017
Predictive	2.13	.036	0.21	.01	.18	0.94	2.73×10^{3}	.036
Short vs. 0								
Causal	2.73	.008	0.27	.02	.13	3.60	3.17×10^{2}	.010
Frequency	0.36	.718	0.04	-.03	.04	0.12	2.02×10^{3}	.535
Predictive	-2.24	.028	-0.22	-.16	-.01	1.17	1.48×10^{3}	.031
Long vs. Short								
Causal	-1.72	.089	-0.25	-.18	.01	0.45	6.44×10^{2}	.068
Frequency	-1.97	.052	-0.26	-.10	.00	0.70	1.87×10^{3}	.050
Predictive	-3.60	<.001	-0.42	-.28	-.08	4.22×10^{1}	1.57×10^{3}	.001

Note. Long vs. 0 presents statistical tests to see if the mean or median is different from zero. The long vs. short comparison subtracts long from short, so positive numbers mean that short is more positive and negative means that short is more negative.

3.2 Generative and Preventive Datasets

Participants clearly learned the generative and preventive relations; the judgments were positive for the generative dataset and negative for the preventive dataset. The statistical tests against zero for both the short and long timeframe were significant and BFs were very strong. Furthermore, the tests comparing the short vs. long timeframe were all non-significant, and most of the BFs were roughly around .12 (about 1:8 in favor of the null hypothesis) and one was 0.19 (about 1:5 in favor of the null). In sum, participants exhibited nearly equivalent ability to learn positive and negative relations in short and long timeframes.

3.3 Illusory Correlation Datasets

3.3.1 A-Cell dataset. For the causal strength and predictive strength measures, the judgments for the A-cell datasets were significantly positive for both the short and long timeframe. The BFs were quite strong, especially for the causal strength judgments, suggesting that participants exhibited illusory correlations. The frequency strength judgments were also positive but the BFs were weaker, especially in the long timeframe, suggesting that different ways to measure illusory correlations may make a difference.

Most importantly, however, the comparisons between short and long timeframes were non-significant for each of the three measures, and the BFs were in the range of .11-.14 (roughly 1:9 or 1:7 in favor of the null). Thus, participants exhibited similar patterns of illusory correlations in the short and long timeframe conditions.

3.3.2 Outcome density. The judgments for the outcome density dataset were somewhat less clear. Outcome density datasets often produce positive illusory correlations, though typically not as strongly as A-cell datasets (Blanco et al., 2013). In the long timeframe condition, there were positive illusory correlations for each measure according to p-values, however, the BFs varied greatly. In the short timeframe condition, some of the measures revealed positive illusory correlations and some were actually negative according to significance testing. The BFs were not at all convincing.

When comparing the short vs. the long timeframe conditions, the causal strength and frequency strength measures were not significantly different (or marginally significant) and the BFs were weak. However, the predictive strength measure was significantly different; the long

timeframe condition produced a positive illusory correlation, but the short timeframe condition produced a negative illusory correlation (inconsistent with prior rapid trial-by-trial studies, e.g., Blanco et al., 2013). For this reason, and since there was only one significant and one marginal difference between long and short timeframes, we believe that these differences between short vs. long timeframe should be interpreted skeptically.

3.4 Judgments During Learning

Figure 3 shows learning curves using the three dependent measures. The causal strength and frequency strength measures are from participants' interim judgments during learning. For predictive strength, participants predicted the presence or absence of the effect on each trial. In Fig. 3, we calculated predictive strength using the predictions from Trials 1-8 for the Trial 8 measure, 9-16 for the Trial 16 measure, and 17-24 for the Trial 24 measure. All three measures have some missing data. There is a bit of missing data for the interim causal strength judgments due to server errors. The frequency and predictive strength have considerable missing data due to being impossible to calculate with a divide by zero error, especially early on.

We did not conduct quantitative analyses for these measures out of concern for inflated Type I error due to the large number of potential comparisons. Qualitatively, the three measures of strength over time are quite similar between the short timeframe and long timeframe conditions. Fig. 3 also shows that substantial learning had already occurred by Trial 8 for causal strength and predictive strength: the generative and preventive conditions were already separated. The frequency strength panel suggests that participants may have had a bit of difficulty remembering the prior evidence at Trial 4 in the long timeframe condition; the judgments were more extreme for the generative and preventive datasets in the short than long timeframe. However, by Trial 12, frequency strength in the short and long timeframes look similar.

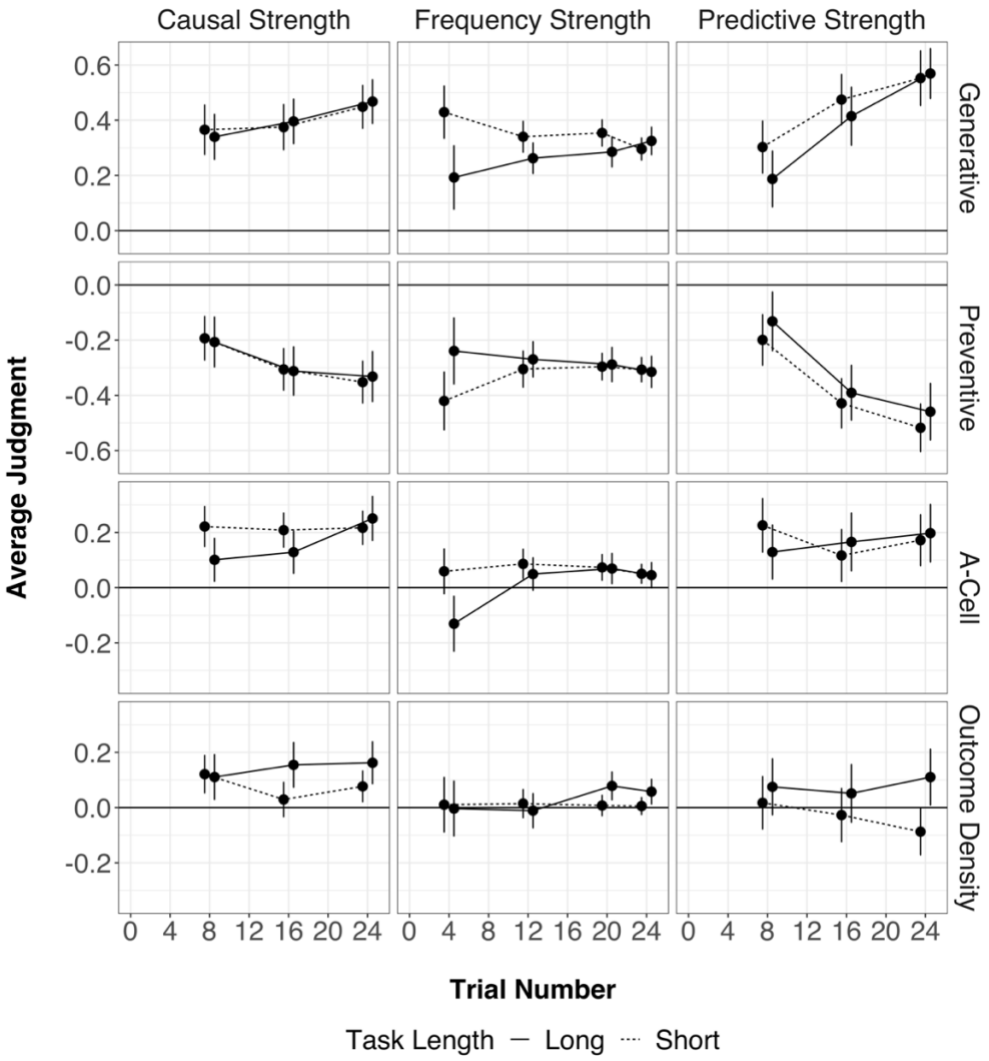

Figure 3. Measures of strength over time for each dataset in the short and long timeframe conditions. The long and short timeframes are separated a bit on the X axis to reduce over-plotting. Error bars represent 95% CI of the mean. Note that frequency strength was measured at different times than causal strength. We reported predictive strength for the same times as causal strength.

3.5 Episodic Memories

In the episodic memory test, participants were shown an image from the study and recalled whether the cause and effect were present or absent, so chance performance for getting both correct was .25. We analyzed all four dataset conditions together. For each participant, we calculated an accuracy score out of the 24 trials for both the short and long timeframe conditions. One participant was excluded from analyses due to a programming error. Thus, these analyses include 412 participants.

Two-sided one-sample t-tests revealed that participants' episodic memories were significantly better than chance in both the short (*M* = .34, *SD* = .12) and long (*M* = .29, *SD* = .11) timeframes (Table 6), though still poor in an absolute sense. The difference between accuracy scores in the short and long timeframe conditions was significant, suggesting that participants had stronger episodic memories in the standard rapid trial-by-trial paradigm.

Table 6. Results for episodic memories collapsed across all datasets. N=412.

	t-test					Bayes Factor
Comparison	*t*	*p*	*d*	*CI lower*	*CI upper*	*BF*
Long vs. 0.25	8.42	<.001	0.41	.28	.30	7.30×10^{12}
Short vs. 0.25	15.18	<.001	0.75	.33	.35	1.68×10^{38}
Short vs. Long*	6.09	<.001	0.38	.03	.06	1.06×10^{6}

Note. *This comparison subtracts long from short, so positive numbers mean that short is more accurate.

4 Discussion

Although standard trial-by-trial paradigms present cue-outcome trials in rapid succession, most real-world experiences require learning from events that are spread out over longer periods of time. We found that when participants observed one trial per day for 24 days, they were capable of learning contingent cause-effect relations, but they exhibited illusory correlations for non-contingent datasets. Critically, we found few differences between the short and long timeframe conditions, with most Bayes factors roughly 8 to 1 in support of the null. We observed some inconsistencies for the outcome density dataset, but these were due to judgments in the short timeframe condition that were inconsistent with prior research.

From a practical perspective, this research suggests that in our everyday lives, humans may be able to detect simple cause-effect relations when they exist; however, we will likely also infer illusory correlations.

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

From a theoretical perspective, we find it striking that there were so few differences in learning across the short and long timeframe conditions. The robust learning in the long timeframe condition is surprising considering that participants completed this condition outside of the lab and likely participated with many distractions and interruptions, as would be expected when learning from experience in everyday life.

4.1 Implications for Theories of Causal Learning

4.1.1 Sequential updating models. Though studies have manipulated the inter-trial-interval (ITI) within human (Msetfi et al., 2005) and animal learning paradigms (Carranza-Jasso et al., 2014; Holland & Morell, 1996), the maximum ITIs have been on the order of minutes. For this reason, empirical research does not provide much guidance about the impact of ITI on learning if the trials are spaced out once per day.

Theoretically, sequential updating models, including reinforcement-learning and associative learning theories (e.g., Rescorla & Wagner, 1972), have very minimal memory demands, which could facilitate learning over long timeframes. Our failure to find differences between short and long timeframes is generally consistent with these classes of models.

4.1.2 Rule-Based models. Rule-based models are the dominant models of how people assess causal relations (Cheng, 1997; Griffiths & Tenenbaum, 2005; Hattori & Oaksford, 2007). These models assume that individuals judge the strength of causal relations from remembered frequencies of the four types of events (each combination of the cue and outcome as present/absent). Therefore, accurate judgments would require accurate tallies of the four types of events.

We found that there were few differences between the causal strength judgments across the short vs. long timeframes, and also that there were few differences between the frequency judgments (tallies). It is possible that learning four tallies is simple enough that it can be performed robustly over long timeframes and that a rule-based approach was used to estimate causal strength. Another possibility is that participants engaged in a sequential updating learning process and recreated the tallies from learned associations.

4.1.3 Uneven cell weighting in attention and memory. People focus most on trials when the cause and effect are present and least on trials when the cause and effect are absent (e.g.,

Kao & Wasserman, 1993). Uneven cell weighting is one of the primary explanations for illusory correlations and is also exaggerated under increased cognitive load (Kao & Wasserman, 1993; Shaklee & Mims, 1982). If spreading out the events over time further taxes memory and attention, by analogy, one might expect a stronger A-cell bias in long timeframes, which would lead to more positive judgments in all four long timeframe conditions. Although we did find illusory correlations, we did not find that the judgments were more positive in the long timeframe condition compared to the short.

4.1.4 Working memory vs. reinforcement learning as alternative learning systems. Recently there have been some advances in research attempting to disentangle the working memory (WM) and reinforcement learning (RL) systems in an instrumental reinforcement learning paradigm (A. G. Collins, 2018; A. G. Collins & Frank, 2012). A key finding was that when participants had to learn which action to take for each of 3 or 6 cues, they were more successful with 3 cues if tested immediately after learning and more successful with 6 cues if tested after a delay (when WM could not be used). This finding and additional modeling suggested that WM and RL both operate in experience-based learning tasks, and to some extent the learner can dynamically trade-off between the two. In the current study, which had only one cue, presumably WM was strongly engaged in the short timeframe. Because participants could not use WM in the long timeframe, either RL or another form of long-term form of memory must have been used. Conducting similar studies in a long timeframe may help reveal whether similar tradeoffs exist and the potential for multiple memory systems in long timeframe learning.

4.2 Implications for Broader Research

4.2.1 Episodic memory. Recently, the possibility has been raised that participants in trial-by-trial learning tasks store and use episodic memories of learning events (Bornstein et al., 2017; Bornstein & Norman, 2017) as opposed to learning an association between the cause and effect or tallies of the four events. We found that the accuracy of participants' episodic memories (29-34%) was significantly but only slightly above chance (25%), and therefore cannot explain the accurate causal judgments observed in the generative and preventive conditions. Furthermore, there were significant timeframe differences for episodic memories but few differences for

causal judgments. Thus, our results suggest that participants were probably not primarily using episodic memories for assessing the strength of the cause-effect relation or for making predictions of the effect from the cause. However, it is possible that participants remembered a few recent learning experiences that had an impact on their assessments.

4.2.2 Timescales of learning. In associative learning, there is a debate about "timescale independence" or "invariance" (Brown et al., 2007; Gallistel & Gibbon, 2000; Kello et al., 2010), in which learning phenomena tend to replicate if the sequence is stretched or compressed. In memory, there are debates about the similarities and differences between short vs. long-term memory (Cowan, 2008) and whether memories across short and long timescales can be modeled with the same forgetting curves (Averell & Heathcote, 2011; Wixted & Ebbesen, 1991). These debates are complex and technical, and though the current study was not designed specifically to address them, the lack of differences speaks to the similarity of learning over short and long timeframes.

4.2.3 Personal semantics. A relatively new topic in memory research is the idea of 'personal semantics' - declarative memories about the self that fall in-between episodic memories of events and semantic memories of facts. Many autobiographical memories involve repeated events (Renoult et al., 2012, 2016). The sorts of memories studied in the current research take autobiographical memories a step further because they involve memories of repeated events involving both a cause (e.g., "I often drinking coffee") and effect (e.g., "I often sleep poorly") and memories about the relationship between the two (e.g., "Drinking coffee impairs my sleep"). Insofar as repeated events are believed to be fundamental to autobiographical memories and beliefs about oneself, the current paradigm is important because it provides a way to empirically study the accuracy of repeated autobiographical memories in a more realistic timeframe.

4.3 Limitations and Future Directions

The current research was intended as a first step towards generalizing the standard short timeframe trial-by-trial paradigm into a long timeframe. In order to make the long timeframe highly comparable to the short timeframe, we intentionally sought a compromise between internal and external validity. Specifically, we manipulated the length of time between the trials

and kept everything else as similar to prior studies as possible; we felt that this was the first most important step towards understanding learning from experience in long timeframes.

We are already modifying the paradigm to test causal learning in progressively more authentic ways. In particular, we are studying whether people can learn more complex cause-effect relations, such as when there are two causes instead of just one, and when there are considerable delays between the cause and the effect. We are also studying situations in which a cause is not used for a period of time and then starts to get used for a period of time (e.g., initially not using a medication and then starting to use a medication for multiple days in a row), rather than the random order of the cause being present or absent in the current study; this alternative task may be harder because it could require remembering back to the prior period of time when the cause was not used.

However, we acknowledge that the long timeframe paradigm we used is still artificial in a number of different ways, some of which we mention here. We hope that this list encourages other researchers from a variety of sub-fields to take up these questions.

4.3.1 Explicit vs. implicit learning. One limitation with almost every causal learning paradigm, including our smartphone paradigm, is that participants are aware of the goal – to learn causal relations – and the smartphone task likely triggers engagement with this goal each day. In the real world, there may be situations in which one has a strong goal to learn and assess a causal relation, for example, when eliminating a food from one's diet to see if it makes them feel better, or when starting a medicine. However, there are also many situations in which one does not initially have a goal, such as encountering a problem (e.g., poor sleep, feeling sick) and then trying to figure out what may have caused the problem. Or there may even be situations that are entirely implicit. Our smartphone paradigm best simulates the most explicit form of learning.

4.3.2 The phone as a protected learning context. Another challenge with using a smartphone to present the materials to participants is that it may serve as its own unique learning context. One of the challenges in learning is the credit assignment problem – identifying which of a practically infinite number of potential causes actually influence an effect (Denniston et al., 2003; Gallistel et al., 2019; Noonan et al., 2017; Staddon & Zhang, 2016;

Sutton, 1984). Perhaps one limitation of the current study is that participants are able to ignore anything happening in their normal lives, and only focus on the single cause and single effect presented on the phone. Thus, our study does not address the credit assignment problem; this is something we are currently studying. At the same time, we note that almost all short timeframe studies are explicitly designed with novel stimuli and a novel context (the lab, or the experimental application on the computer) to accomplish the same goal of protecting the study from outside influences. Furthermore, if it is true that the phone serves as a protected environment for learning in the long timeframe, and that within this environment learning occurs unimpeded relative to the short timeframe, in our opinion, this is an important finding in its own right. In theory, it should be entirely possible for events outside the phone to affect learning that occurs within the context of the phone to some extent (e.g., through interference or simply distraction). Furthermore, the length of time between trials should impact memory decay regardless of the phone context. The fact that learning was so similar in the long and short timeframes highlights the important role of attention at facilitating learning and it also demonstrates that the length of the task does not inherently impair learning.

4.3.3 Causal strength vs. causal structure. One of the most major developments in research on causation in the past two decades has been the shift from only focusing how people learn causal strength towards how people learn the causal structure among multiple variables (Bramley et al., 2018; Coenen et al., 2015; Davis et al., 2020; Rothe et al., 2018; Rottman & Keil, 2012; Steyvers et al., 2003). One reason that this is vitally important is because it is helps to distinguish truly causal vs. associative representations (though see Fernando (2013) for an associative learning model of causal structure). Whereas associative representations have very low memory demands, inferring causal structures with causal representations requires considerably more memory and reasoning demands. Thus, another interesting question for future research is the extent to which people can learn causal structure in more realistic contexts.

4.3.5 Definition of a trial. Another limitation of this paradigm, and trial-by-trial paradigms generally, is that the meaning of a 'trial' is very hard to define let alone justify from a naturalistic learning perspective (Gallistel, 2021; Gallistel & Gibbon, 2000). For example, one

way in which the 'trial' paradigm is artificial is that absences are more salient than usual; in the real world, not taking medicine and not experiencing pain are typically less salient than taking medicine and feeling pain. In most causal learning paradigms, however, the absence and presence of the cause are explicitly stated and therefore made artificially salient. Another related concern is that many real-world phenomena are not punctate events that happen at a brief moment in time, but instead may persist or fluctuate in intensity over periods and grow or recede in awareness over time (e.g., pain, mood; Davis et al., 2020; Soo & Rottman, 2015. Studying causation with other paradigms over long timeframes will help make the research more realistic.

4.3.4 Repeated judgments. One aspect of the current design is that on certain learning trials, participants were prompted to make causal judgments or to report their memories. We did this in order to have interim measures of learning aside from the predictions. Prior research has varied widely in the practices used. The majority of studies on causal learning simply ask for judgments at the end of learning. Some have asked for judgments a few times during learning, and a few have asked for judgments on every trial (Soo & Rottman, 2015; Van Hamme & Wasserman, 1994). There is some evidence that asking for judgments repeatedly may lead participants to report only what has happened on the most recent experiences since the last judgment (D. J. Collins & Shanks, 2002; Hogarth & Einhorn, 1992; Marsh & Ahn, 2006). It is possible that asking for interim judgments helped improve learning or modified learning in some way. That said, given that interim judgments were asked for both the long and short timeframe conditions, it seems unlikely that such judgments are responsible for the essentially equal performance in the two conditions; it would have needed to perfectly balance out any other differences between the short vs. long timeframe conditions.

4.3.6 Within vs. between subjects design. A weakness of our study is that it used an atypical within-subjects design, with the long timeframe task occurring in-between the short timeframe tasks on the first and last day. We implemented the within-subjects comparison to increase power and because a between subjects design is very expensive (participants in the short and long conditions would need to be paid the same amount). That said, researchers who want to study long timeframe learning may consider doing a standard counterbalanced design,

or may consider only studying the long timeframe condition instead of comparing long vs. short timeframe learning.

4.4 Potential for Mobile Phones and Wearables for Ecological Momentary Experiments

We believe that mobile phones, smart watches, and other wearable technologies provide valuable but largely untapped opportunities for researchers to conduct experiments that are embedded in everyday life. For many years various technologies have facilitated the collection of survey data at different times of day using ecological momentary assessment (EMA), experience sampling, and intensive longitudinal data in order to prevent recall bias and other problems with retrospective reports (e.g., Shiffman et al., 2008). The omnipresence of smartphones is now facilitating ecological momentary interventions (EMI) – interventions meant to prompt behavior change around mental and behavioral health (Heron & Smyth, 2010; Marcolino et al., 2018; Nahum-Shani et al., 2018; see also just in time active interventions, JITAI, and mobile health, mhealth).

The current research is a departure from these methods in that it involves embedding what would otherwise be a normal psychology experiment, with high degrees of manipulation and control of stimuli, into participants' lives; we call this approach 'ecological momentary experiments' (EME). Though such experiments are hard to program and conduct, we believe that they are valuable because a wide variety of learning phenomena that are typically studied in short timeframe studies translate best to phenomena that play out over considerably longer periods of time in real life. Thus, we believe that this EME paradigm can be adapted to studying many phenomena studied by cognitive scientists (e.g., probability learning, second language learning, memory, decision-making, sleep, and others).

In our study, we used a custom-designed program based on our PsychCloud.org framework and hosted on Google's app engine. Recently a number of other tools have become available for both presenting stimuli and asking questions (Mack et al., 2019; Shevchenko et al., 2021; Stieger et al., 2018). We hope that as these technologies become more powerful and easier to use, researchers can exploit the possibilities that they open up.

4.5 Conclusions

The current research is an important step towards generalizing existing learning paradigms to more real-world settings; it shows that people have similar strengths and weaknesses when learning simple cause-effect relations in long timeframes as they do in short timeframes. However, we also raised a number of limitations with the study – there are still many ways in which real world causal learning differs from the current paradigm and there are also many different sorts of real-world causal learning situations. Additionally, our participants were mostly college students, but older adults with age-related working memory decline exhibit less accurate causal learning (Mutter et al., 2009; Mutter & Plumlee, 2009); it is uncertain how accurately older adults can perform causal learning over long timeframes. Given how important causal learning and statistical learning are for so many cognitive functions, additional research is needed to understand the accuracy of these abilities in real-world environments.

References

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.

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.

Bott, F. M., & Meiser, T. (2020). Pseudocontingency inference and choice: The role of information sampling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advanced Online Publication. http://dx.doi.org/10.1037/xlm0000840

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

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

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.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*(3), 271–280.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Collins, A. G. (2018). The tortoise and the hare: Interactions between reinforcement learning and working memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035.

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

Cowan, N. (2008). What are the differences between long timeframe, short timeframe, and working memory? *Processes in Brain Research*, *169*, 323–338.

Davis, Z. J., Bramley, N. R., & Rehder, B. (2020). Causal structure learning in continuous systems. *Frontiers in Psychology*, *11*, 244.

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.

Denniston, J. C., Savastano, H. I., Blaisdell, A. P., & Miller, R. R. (2003). Cue competition as a retrieval deficit. *Learning and Motivation*, *34*(1), 1–31. https://doi.org/10.1016/S0023-9690(02)00505-2

Eder, A. B., Fiedler, K., & Hamm-Eder, S. (2011). Illusory correlations revisited: The role of pseudocontingencies and working-memory capacity. *The Quarterly Journal of Experimental Psychology*, *64*(3), 517–532.

Fernando, C. (2013). From blickets to synapses: Inferring temporal causal networks by observation. *Cognitive Science*, *37*(8), 1426–1470.

Gallistel, C. R. (2021). Robert Rescorla: Time, Information and Contingency. *Revista de Historia de La Psicología*, *41*(1), 7–21. https://doi.org/10.5093/rhp2021a3

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

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

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

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*(4), 392–407.

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.

Herek, G. M. (2010). Sexual orientation differences as deficits: Science and stigma in the history of American psychology. *Perspectives on Psychological Science*, *5*(6), 693–699.

Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, *15*(1), 1–39.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55.

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.

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*, 138–161.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the “same”? Coherent generalization across contexts. *Psychological Science*, *18*(11), 1014–1021.

Mack, C. C., Harding, M., Davies, N., & Ward, G. (2019). RECAPP-XPR: A smartphone application for presenting and recalling experimentally controlled stimuli over longer timescales. *Behavior Research Methods*, *51*(4), 1804–1823.

Marcolino, M. S., Oliveira, J. A. Q., D'Agostino, M., Ribeiro, A. L., Alkmim, M. B. M., & Novillo-Ortiz, D. (2018). The impact of mHealth interventions: Systematic review of systematic reviews. *JMIR MHealth and UHealth*, *6*(1), e23.

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

Matute, H., Yarritu, I., & Vadillo, M. A. (2011). Illusions of causality at the heart of pseudoscience. *British Journal of Psychology*, *102*, 392–405.

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

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.

Mullen, B., & Johnson, C. (1990). Distinctiveness-based illusory correlations and stereotyping: A meta-analytic integration. *British Journal of Social Psychology*, *29*, 11–28.

Mutter, S. A., DeCaro, M. S., & Plumlee, L. F. (2009). The role of contingency and contiguity in young and older adults' causal learning. *Journal of Gerontology: Series B*, *64*(3), 315–323.

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.

Mutter, S. A., & Plumlee, L. F. (2009). Aging and integration of contingency evidence in causal judgment. *Psychology and Aging*, *24*(4), 916–926.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, *52*(6), 446–462.

Noonan, M. P., Chau, B. K. H., Rushworth, M. F. S., & Fellows, L. K. (2017). Contrasting effects of medial and lateral orbitofrontal cortex lesions on credit assignment and decision-making in humans. *The Journal of Neuroscience*, *37*(29), 7023–7035. https://doi.org/10.1523/JNEUROSCI.0692-17.2017

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.

Redelmeier, D. A., & Tversky, A. (1996). On the belief that arthritis pain is related to the weather. *Proceedings of the National Academy of Sciences*, *93*(7), 2895–2896.

Renoult, L., Davidson, P. S., Palombo, D. J., Moscovitch, M., & Levine, B. (2012). Personal semantics: At the crossroads of semantic and episodic memory. *Trends in Cognitive Sciences*, *16*(11), 550–558.

Renoult, L., Tanguay, A., Beaudry, M., Tavakoli, P., Rabipour, S., Campbell, K., Moscovitch, M., Levine, B., & Davidson, P. S. (2016). Personal semantics: Is it distinct from episodic and semantic memory? An electrophysiological study of memory for autobiographical facts and repeated events in honor of Shlomo Bentin. *Neuropsychologia*, *83*, 242–256.

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.

Rothe, A., Deverett, B., Mayrhofer, R., & Kemp, C. (2018). Successful structure learning from observational data. *Cognition*, *179*, 266–297.

Rottman, B. M. (2019). *Overview*. PsychCloud. www.psychcloud.org

Rottman, B. M., & Keil, F. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*(1–2), 93–125.

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.

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.

Shevchenko, Y., Kuhlmann, T., & Reips, U.-D. (2021). Samply: A user-friendly smartphone app and web-based means of scheduling and sending mobile notifications for experience-sampling research. *Behavior Research Methods*, 1–21.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32.

Soo, K., & Rottman, B. M. (2015). *Elemental Causal Learning from Transitions* (R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi, Eds.). Austin, TX: Cognitive Science Society.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science*, *40*(3), 305–313.

Staddon, J. E., & Zhang, Y. (2016). On the assignment-of-credit problem in operant learning. In M. L. Commons, S. Grossberg, & J. E. Staddon (Eds.), *Quantitative analyses of behavior series. Neural network models of conditioning and action*. Hillsdale, NJ: Erlbaum.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Stieger, S., Lewetz, D., & Reips, U.-D. (2018). Can smartphones be used to bring computer-based tasks from the lab to the field? A mobile experience-sampling method study about the pace of life. *Behavior Research Methods*, *50*(6), 2267–2275.

Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning* [Doctoral dissertation, University of Massachussets]. http://scholarworks.umass.edu/dissertations/AAI8410337

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, *25*(2), 127–151.

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.

Appendix

Table A1. Tests of order effects

			Causal Strength			Frequency Strength			Predictive Strength		
Dataset	Length	*df*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*
Generative	Short	97	0.91	.364	0.18	0.47	.643	0.09	0.29	.774	0.06
	Long	97	0.46	.646	0.09	0.14	.887	0.03	-1.33	.185	-0.27
Preventive	Short	102	1.01	.317	0.20	0.71	.477	0.14	2.37	.020*	0.46
	Long	102	0.75	.457	0.15	1.18	.241	0.23	0.18	.856	0.04
A-Cell	Short	104	1.66	.101	0.32	1.54	.126	0.30	1.43	.156	0.28
	Long	104	1.59	.116	0.31	-1.30	.196	-0.25	-0.20	.842	-0.04
Outcome	Short	102	0.75	.458	0.15	-0.44	.660	-0.09	-0.27	.790	-0.05
Density	Long	102	-0.25	.803	-0.05	-1.34	.183	-0.26	-0.67	.506	-0.13

Note. *$p<.05$. Positive numbers indicate more positive judgments if the task was completed first than if the task was completed second.