

EMPIRICAL ARTICLE

# Causal learning with interrupted time series data

Yiwen Zhang  and Benjamin M. Rottman 

Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA

**Corresponding author:** Yiwen Zhang; Email: [yiwenzhang@pitt.edu](mailto:yiwenzhang@pitt.edu)

**Received:** 18 November 2022; **Revised:** 11 July 2023; **Accepted:** 17 July 2023

**Keywords:** causality; time series; illusory causation; memory

## Abstract

People often test changes to see if the change is producing the desired result (e.g., does taking an antidepressant improve my mood, or does keeping to a consistent schedule reduce a child's tantrums?). Despite the prevalence of such decisions in everyday life, it is unknown how well people can assess whether the change has influenced the result. According to interrupted time series analysis (ITSA), doing so involves assessing whether there has been a change to the mean ('level') or slope of the outcome, after versus before the change. Making this assessment could be hard for multiple reasons. First, people may have difficulty understanding the need to control the slope prior to the change. Additionally, one may need to remember events that occurred prior to the change, which may be a long time ago. In Experiments 1 and 2, we tested how well people can judge causality in 9 ITSA situations across 4 presentation formats in which participants were presented with the data simultaneously or in quick succession. We also explored individual differences. In Experiment 3, we tested how well people can judge causality when the events were spaced out once per day, mimicking a more realistic timeframe of how people make changes in their lives. We found that participants were able to learn accurate causal relations when there is a zero pre-intervention slope in the time series but had difficulty controlling for nonzero pre-intervention slopes. We discuss these results in terms of 2 heuristics that people might use.

## 1. Introduction

Many situations in life, in both formal and informal decision-making, involve tracking an individual before and after a change to assess whether the change appears to have produced a desired outcome. These sorts of situations can be considered cases of an 'interrupted time series', in which an intervention 'interrupts' a time series and may produce a change in the level or slope of the outcome.

The goal of this article is to test if people make judgments that line up with formal interrupted time series analysis (ITSA) when assessing causal relations in a variety of single-subject situations. We tested judgments when the experiences were presented in multiple ways: simultaneously, such as reading a graph; sequentially in short succession like in many typical cognitive psychology paradigms; and most importantly, once per day over 14 days to understand the accuracy of causal judgments in a situation more akin to everyday life.

In the introduction, we first introduce interrupted time series designs and the statistical theory behind ITSA. Then we talk about previous empirical research on how people reason from time series data in general, and interrupted time series data in particular, and models that could explain how people reason about interrupted time series data. Afterward, we talk about research on how different presentation

formats may lead to different judgments for time series data, and most importantly, how people may reason about interrupted time series data when presented in a more naturalistic environment when spread out once a day over many days. Finally, we outline the goals of the current studies.

### ***1.1. Randomized controlled trials versus interrupted time series designs***

In many situations that involve making a decision, it would be ideal to conduct a randomized controlled experiment to provide guidance for the decision. However, there are at least 2 challenges with randomized controlled trials that prevent them from being useful in many common situations, and in these situations, an interrupted time series design is often the best option available.

First, conducting a randomized controlled trial is often impractical. For example, in many economic situations, a policy change is implemented and one may wish to assess whether the change was efficacious. However, it is usually politically impossible or extremely challenging to conduct a randomized controlled trial of economic policies because it is usually impossible to randomize territories to different policies.

Second, at a much smaller level, when an individual needs to make a decision for themselves or their family, even if there is a randomized controlled trial that provides advice about which option is better on average, randomized controlled trials can only provide guidance about averages. The best option for an individual may deviate from what is best on average. Relatedly, certain questions (e.g., is increasing the minimum wage by a certain amount at a certain time better than not increasing the minimum wage, or is switching a child's school in 10th grade the best thing for them at that time?) are unanswerable with a controlled trial because there is only one of each individual (countries, people); we can never know what would have happened if the counterfactual option had been chosen.

In situations such as these, often the best that can be done is to track an individual (person, country, or other entity) over time, before and after a change has been implemented, to determine if the change is having the desired consequence. Tracking an individual before and after a change is known as an 'interrupted time series' design, and the statistical method used for this design is called 'ITSA' (Bernal et al., 2017).

Interrupted time series designs are used in a variety of circumstances from formal research to informal everyday decision making. First, they are used in formal research, for example, economic research mentioned above. Because there can always be confounds—other factors that have changed at roughly the same time—economists have devised creative ways to use control cases to uncover causal effects in interrupted time series data (The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel, 2021).

Second, interrupted time series designs are used in single-subject studies in which the goal is to assess whether a change is working for an individual, but not to draw broader conclusions about groups. They are especially common in applied behavior analysis (Barlow and Hayes, 1979) and 'elimination diets' or 'challenge trials', which test for food allergies by eliminating and then reintroducing different foods (e.g., Minford et al., 1982). More broadly, interrupted time series is part of the idea of developing 'adaptive interventions' or 'individualized treatment sequences', which are interventions that are dynamically adjusted based on individual patient's needs and treatment responses (Lei et al., 2012).

Third, any situation in which one makes a change and tracks outcomes can be viewed as a sort of informal interrupted study design. Thus, many decisions that people make as part of everyday life to adapt to the environment are essentially a kind of interrupted study design. For example, a large study of adults older than 65 found that in a 1 year period of time, 24% said that they skipped doses of a medication or stopped taking it ('nonadherence') because they believed that it was making them feel worse and or that they did not think that the medication was helping (Wilson et al., 2007). These results attest to the fact that people make decisions about important choices based on their personal interpretation of their experiences from interrupted time series situations; however, it is unclear how well these sorts of decisions are made. Most people in these situations probably do not record their

experiences and instead make decisions from memory. Even if they were to record experiences, would the decisions be better?

In sum, interrupted time series designs are widespread in formal and informal decision-making. In the next section, we discuss formal statistical analysis of interrupted time series data and then talk about prior research on how lay people make judgments about time series data in general and interrupted time series data in particular.

## 1.2. Interrupted time series analysis

ITSA is a statistical procedure that can be used with single ‘subject’ (e.g., any single entity such as a person, group, or economy) data to assess the potential impact of an intervention on an outcome (Bernal et al., 2017; Hartmann et al., 1980). Interrupted time series refers to a time series in which the putative cause is changed once (e.g., from absent to present or one level to another), which is sometimes called an ‘AB design’. In ‘reversal’ time series designs, the putative cause may change twice (‘ABA’) or three times (‘ABAB’); in the current research, we focused on the simplest AB design.

A simple ITSA<sup>1</sup> is a regression model which detects if there is a significant change in the slope or in the level of the time series data after the intervention was made, compared to the pre-intervention period. The regression model consists of 3 components which are a pre-intervention slope, a slope change, and a level change. In equation (1), the time variable ( $T$ ) is the time period since the observation started and the intervention variable ( $X$ ) is coded as 0 before the intervention was introduced and after the intervention was introduced. Coefficient  $\beta_1$  captures a change over time, which is also called the pre-intervention slope. Coefficient  $\beta_2$  indicates a level change that occurs after the intervention was introduced compared to before. The interaction coefficient  $\beta_3$  indicates a slope change after the intervention compared to before.

$$Y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 XT. \quad (1)$$

Because ITSA applies to a single entity and often does not include a control, making a causal inference that the intervention caused a change in the slope and or level requires the additional assumption that there are no other confounds that occurred close in time with the intervention. Though of course sensitivity to the possibility of confounds is important, this is not the goal of the current research (see Derringer and Rottman, 2018 for related research). In the current research, we only present a single potential cause, and the cover story being used implies that the interrupted time series intervention is being conducted as part of a single-subject clinical trial, raising the plausibility that any change in level and slope of the outcome are in fact caused by the intervention. Our key question is whether people can draw appropriate conclusions in interrupted time series situations, under the assumption of no confounds.

We investigated how people reason about a variety of datasets that can be generated from the interrupted time series model. In particular, we tested participants’ judgments about cases that include pre-intervention slopes or not, changes to the level after the intervention or not, and changes to the slope after the intervention or not.

## 1.3. Causal learning from time series data and interrupted time series data

Recently there has been a growing trend of research into how people learn about and reason from time series data. By time series, we mean situations in which the state of a variable at one point in time may be related to its state at the prior time, as opposed to situations in which there is no inherent order to the data or the data are randomized so that the order is irrelevant. Some studies have investigated how people learn cause-effect relations from time series data with multiple potential causes (Derringer and

<sup>1</sup>There are various models to analyze interrupted time series data and all the models have their own strengths in analyzing data with different properties, for example, serial dependency, nonstationary, and nonconstant pre-intervention slope.

Rottman, 2018; Lagnado and Speekenbrink, 2010), and multiple studies have investigated how people learn causal structure—which variables cause which others—from different sorts of time series data (Bramley et al., 2017, 2018; Davis et al., 2020; Rehder et al., 2022; Rottman and Keil, 2012; Rottman et al., 2014). The current research focuses on a simpler case in which there are only 2 variables, one of which is a potential cause and the other is a potential effect, so it does not involve learning the structure (direction) of the relation.

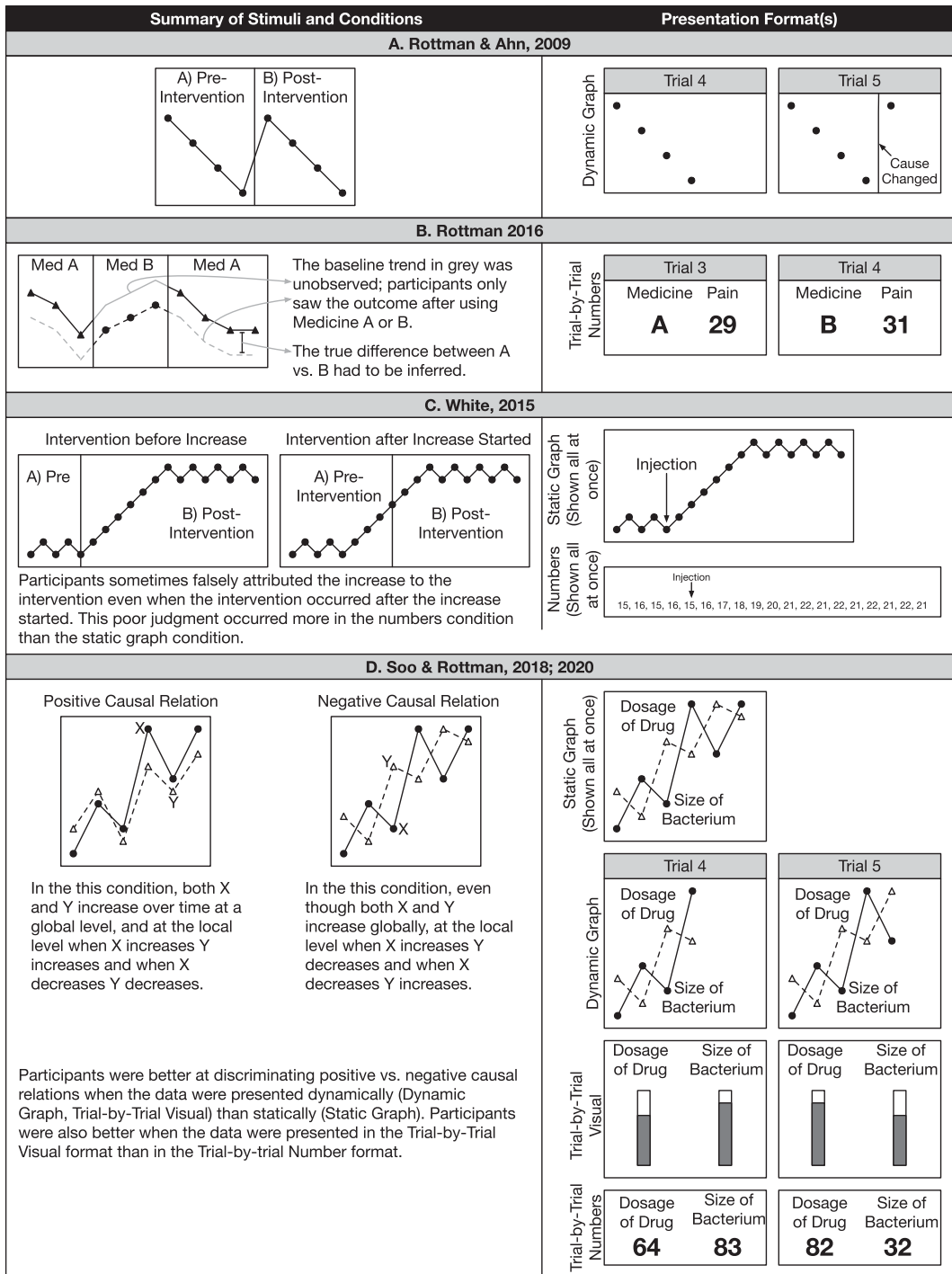
How people assess the causal relation between 2 variables has been studied for decades; however, how people assess the relation between 2 variables in single-subject time series data is a more recent trend. One of the fundamental ideas in research on how people interpret time series data is that people do not just compare the average outcome when the cause is present to the average outcome when the cause is absent, or other similar measures that are appropriate for cross-sectional data (e.g., Cheng, 1997; Griffiths and Tenenbaum, 2005). For example, one of the first studies of this sort investigated how people think about cause-effect relations that experience increasing effectiveness over time (e.g., antidepressants) or decreasing effectiveness (e.g., caffeine; Rottman and Ahn, 2009). In such a situation there can be zero bivariate correlation between a cause (e.g., number of cups of coffee) and an effect (e.g., wakefulness), yet people may still conclude that the cause made a difference, by paying attention to the patterns of the variables over time (Figure 1A). Indeed from an ITSA perspective in the example in Figure 1A, there is a level change.

One challenge in learning causal relations from time series data is that the data often have trends that need to be controlled for in order to uncover the true relationship. For example, Rottman (2016) investigated how people test cause-effect relations when the effect undergoes unpredictable wavy baseline trends (Figure 1B). When making a final judgment of whether the effect was better given different levels of the cause, in addition to comparing the average of the effect for different levels of the cause, they also compared the averages of how the effect changed from one experience to the next (first order derivative). For example, in Figure 1B, the patient used Medicine B on Days 4–6, and on these days the amount of pain tended to increase, but on the other days when the patient was using Medicine A, the pain tended to decrease. Taking change scores is a standard way of controlling for trends in time series analysis (Soo and Rottman, 2018, 2020). More broadly, focusing on change scores also suggests that people may notice changes in slope, which is an important aspect of ITSA.

White (2015, 2017) investigated how people reason about AB scenarios in which the outcome starts out flat, then increases and reaches a plateau (Figure 1C), or reaches a plateau and then comes back down to the initial level. The most surprising finding was that even in situations in which the outcome had already started to increase prior to the intervention (right panel in Figure 1C), participants still tended to conclude that the intervention had some impact on the increase. Whether people conclude that an intervention that occurs after a trend that has already started is another question of our study.

Finally, multiple studies have examined how applied behavior analysts who are trained to interpret single-subject graphs, as well as lay individuals, interpret single-subject interrupted time series graphs (Bishara et al., 2021; Deprospero and Cohen, 1979; Jones et al., 1978). The judgments often did not often agree with the conclusions from ITSA and were relatively unreliable, with only modest inter-subject consistency. High serial dependency (autocorrelation) in the datasets was identified as one reason that the judgments deviated from formal ITSA. When serial dependence was present, people tended to incorrectly conclude that the intervention produced a change in the outcome even when it did not.

In summary, there have been multiple lines of research about how people make judgments from interrupted time series data. Most relevantly to the current research, some of the past research suggests that people may attend to changes in slope (Rottman, 2016) and that people may conclude that an intervention has caused a change in slope or level even if the change occurred before the intervention (White, 2015, 2017). However, these studies did not use the standard ITSA generative process (equation (1)) and instead examined more complex processes like wavy baseline trends or situations without a clear statistical generative mechanism. The studies that have investigated the standard linear interrupted time series generative process with changes in level and slope (e.g., Bishara et al., 2021)



**Figure 1.** Time series datasets used in prior studies.

*Note:* The stimuli and presentation formats are intended to be illustrative, not exact replications.

|                                                                                   |                               | ITSA Coefficients |              |              | Qualitative Model Predictions |     |             |              | Summary of Results |      |        |      |        |      |
|-----------------------------------------------------------------------------------|-------------------------------|-------------------|--------------|--------------|-------------------------------|-----|-------------|--------------|--------------------|------|--------|------|--------|------|
| Condition                                                                         |                               | Pre. Slope        | Level Change | Slope Change | ITSA                          | A-B | Post. Trend | RW           | Exp. 1             |      | Exp. 2 |      | Exp. 3 |      |
| No Influence Conditions                                                           |                               |                   |              |              |                               |     |             |              | C.S.               | F.U. | C.S.   | P.S. | C.S.   | P.S. |
|  | A. Flat                       | 0                 | 0            | 0            | 0                             | 0   | 0           | 0            | 0                  | +    |        |      |        |      |
|  | B. Pre-intervention Slope     | +                 | 0            | 0            | 0                             | +   | +           | +            | +                  | *    | +      | 0    | +      | +    |
| Level Change Conditions                                                           |                               |                   |              |              |                               |     |             |              |                    |      |        |      |        |      |
|  | C. Level Change               | 0                 | +            | 0            | +                             | +   | 0           | +            | +                  | +    | +      | +    | +      | +    |
|  | D. Level Change (Congruent)   | +                 | +            | 0            | +                             | +   | +           | +            | +                  | +    |        |      |        |      |
|  | E. Level Change (Incongruent) | -                 | +            | 0            | +                             | 0   | -           | Close to 0 † | -/0*               | -/0* | 0      | 0    | 0*     | -    |
| Slope Change Conditions                                                           |                               |                   |              |              |                               |     |             |              |                    |      |        |      |        |      |
|  | F. Slope Change               | 0                 | 0            | +            | +                             | +   | +           | +            | +                  | +    | +      | +    | +      | +    |
|  | G. Slope Change (Congruent)   | +                 | 0            | +            | +                             | +   | +           | +            | +                  | +    |        |      |        |      |
|  | H. Slope Change (Incongruent) | -                 | 0            | +            | +                             | 0   | +           | -/0/+ †      | +                  | +    | +      | 0    | +      | +    |
|  | I. Slope Change (Maintain)    | -                 | 0            | +            | +                             | -   | 0           | -            | 0*                 | 0/+* | 0      | 0    | 0      | 0    |

**Figure 2.** The overview of interrupted time series conditions, qualitative model predictions, and summary of results.

Note: A–B = After-minus-Before; CS = causal strength; F.U. = future use; PS = predictive strength. + means higher than 0 and – means lower than 0. Because each cell aggregates results from both frequentist and Bayesian analyses and from multiple conditions, the findings reported in a cell should be viewed as an attempt to capture a summary of the main patterns, but do not imply complete agreement of all the results represented within the cell. † The RW predictions for Conditions E and H depend on the choice of the learning rate. \*There are individual differences in this measure.

largely focused on an additional component of serial dependency. Thus, the goal of the current research was to focus on how people make judgments about the simpler case of changes in level and slope in a linear interrupted time series process.

In particular, we studied judgments about 9 types of datasets (Figure 2). These are divided into 3 groups: no change in level or slope (A and B) which implies no causal influence of the intervention, a change in level (C–E), and a change in slope (F–I). In each of these groups, some of the graphs have a zero pre-intervention slope and some have a pre-intervention slope.

1.4. Theories for judging interrupted time series data

In the current study, we focused on 4 potential theories for how people might make inferences from interrupted time series data. First, participants might use a *process akin to ITSA* which would essentially involve noticing whether or not there is a change in intercept and or slope after versus before the intervention. In the current studies, there was very little noise, so this should not be a hard perceptual judgment at least when the data are presented at once in a graph.

Second, one heuristic is the *After-minus-Before* model proposed by White (2015), which involves subtracting the mean value of the outcome during the pre-intervention period from the mean value during the post-intervention period and does not account for any pre-intervention trends. The After-minus-Before heuristic was able to predict most of the results in White’s experiments (White, 2015, 2017), such as the finding that participants conclude that an intervention has an influence even if the outcome started to change prior to the intervention. Bishara et al. (2021) proposed the same heuristic (they called it the ‘absolute difference heuristic’) and noted that it could account for the finding that participants tended to conclude that the intervention caused a difference in the outcome, even when it



actually did not, due to high serial dependence. However, the After-minus-Before heuristic does not explain all the findings. For example, White (2017) found participants gave lower causal judgments for [12, 14, 16, 18, 20, intervention, 22] than [20, 20, 20, 20, 20, intervention, 22], even though the After-minus-Before model predicts a larger judgment for the first; presumably, the participants noticed the increasing pre-intervention trend and thought that the outcome would continue to increase even without the intervention.

The After-minus-Before heuristic agrees with ITSA in 5 conditions (A, C, D, F, and G in Figure 2). It incorrectly judges a positive causal strength in Condition B because the post-mean is higher than the pre-mean even though there is zero change to the slope or level. It incorrectly judges zero causal strength in Condition E and I because the means of the 2 periods are the same despite changes to the level (E) or slope (I). It also incorrectly judges a negative influence in Condition H due to the lower mean in the post-intervention period, even though there is a positive change in the slope.

Third, another heuristic involves simply focusing on the slope of the *post-intervention trend*. We initially came up with this theory from some of our participants' explanations for judgments in a pretest in which we asked them to explain how they judged the interrupted time series graphs; some participants only mentioned a slope in the post-intervention period and did not mention the pre-intervention period at all. The idea is that a positive (negative) post-intervention slope is interpreted as evidence that the intervention increases (decreases) the outcome, no matter what trend occurred in the pre-intervention period. This sort of inference is clearly nonrational but it could be understood from a belief that something must be responsible for a change in the outcome, so if the outcome increases while the cause is present, then the cause could be responsible for this increase.

The Post-intervention Trend heuristic agrees with ITSA in 5 conditions (A, D, F, G, and H in Figure 2). Similar to the After-minus-Before heuristic, it predicts a positive (or negative) causal strength for Condition B because the outcome continues to go up (or down) in the post-intervention period. Unlike the After-minus-Before heuristic, it predicts zero influence in Conditions C and H and a negative influence in Condition E.

Fourth, many researchers have considered using models of associative learning such as the *Rescorla and Wagner (1972) model* (RW) as a model of causal learning (e.g., Matute et al., 2019; Shanks, 2007). At asymptote, RW can detect average changes in an outcome given the presence of a cue similar to linear regression (Danks, 2003). However, the asymptotic guarantees only work under randomized trial orders; the AB situation being investigated here does not use a randomized trial order.

We present simulations of RW for all 9 AB interrupted time series we study in Appendix A. For 7 out of the 9 conditions, RW makes the same qualitative predictions as the After-minus-Before model. For Condition E, the predictions are close to 0, but not exactly 0, and depend a bit on the learning rate parameter. For Condition H, the final associative strength was somewhat positive for high learning rates and slightly negative for low learning rates. Because all 3 other theories make definitive predictions that do not depend on parameters, and because the predictions of RW are so similar to the After-minus-Before model, we mainly focus on the other 3 models and come back to RW in the general discussion.

In sum, by testing 9 different conditions and comparing them to the qualitative conditions of these models we will assess if any of the models explain participants' behavior better than others. That said, the models often make qualitatively similar predictions for many of the conditions; we did not approach this project with the primary goal of testing which model would explain the results best as we assumed that it was likely that none of the models would explain all the findings.

### 1.5. Presentation formats

When studying causal learning with time series data, researchers have used various formats to present data to participants. Some studies have investigated how people reason with time series graphs (Caddick and Rottman, 2019; Deprospero and Cohen, 1979; Jones et al., 1978). Others have investigated time series data presented as a list of numbers (White, 2015). And some studies have investigated time series data using a trial-by-trial paradigm in which the data is shown one at a time (Rottman and Ahn,

2009; Soo and Rottman, 2018) or a continuous paradigm in which the quantity of stimuli increases or decreases over time (Gong and Bramley, 2022), similar to how an individual experiences events unfolding over time.

White (2015, 2017) investigated how participants interpreted AB designs when presented as a list of numbers versus in a time series line graph (Figure 1C). In the situation in which the increase in the outcome started before the intervention, so the intervention could not be the cause White found that fewer participants incorrectly inferred an influence of the intervention in the line graph condition. It seemed that the graph format helped participants detect the true causal relationship.

Soo and Rottman (2018, 2020) also investigated how various formats affect causal inferences from time series data. However, unlike most of the previous studies discussed so far, which used a binary (present vs. absent) cause and could be considered as an AB or ABA situation, these studies used a continuous cause and effect (Figure 1D). This involved situations in which 2 variables could exhibit strong increasing or decreasing trends over time, and one also caused the other to increase or decrease. In addition to being influenced by the overall correlation, participants also noticed these more local relations and used them for inferring the causal relationship. Soo and Rottman (2018) also found that participants were better at accounting for trends and noticing the local changes when a sequential trial-by-trial presentation used shapes that changed in size than when it involved a number that changed from one trial to the next (compare Trial-by-Trial Visual vs. Trial-by-Trial Numbers in Figure 1D). Presumably, it is easier to visually notice changes in shape and size whereas numbers need to be mentally converted into magnitudes before being compared. Soo and Rottman (2020) further found that dynamic presentations (when the experiences were presented trial-by-trial or a line graph that was gradually revealed) helped people more accurately learn the causal relationship by focusing on local changes in the cause and effect from one trial to the next. In contrast, a static line graph like in Figure 1 led them to focus more on the global correlation (compare Trial-by-Trial Visual and Dynamic Graph vs. Static Graph in Figure 1D).

In sum, the existing research suggests that dynamic presentations and presentations that make use of time series graphs rather than lists of numbers tend to allow people to notice trends better, which tends to lead to better causal judgments. However, none of the studies that compared different presentation formats investigated the simple AB interrupted time series situation with potential changes to the slope and or level. Thus, one of the goals of the current study was to test if similar findings hold in the AB interrupted design case.

On the one hand, these studies do seem to be pointing in a fairly consistent direction that dynamic and visual presentations are better for causal inference than static and numeric presentations. On the other hand, it may be the case that in simple AB-interrupted time series situations the trends are easy enough to notice so that there are no considerable differences between conditions. In particular, in Soo and Rottman's studies (2018, 2020), it was vital for participants to notice if a change in the cause was associated with a change in the effect at a local level and *not* to focus on the global trends in each variable but instead to control for these global trends. In contrast, in the simple AB situation studied here, the vital task was to focus on the global trends in the A and B periods and to compare these 2. In sum, even if dynamic and visual presentations are especially helpful for getting people to notice local (from one trial to the next) changes, this may be less important for interpreting AB designs, which are more about noticing global patterns.

To test this question, we compared a static time-series graph with a dynamic time-series graph that was sequentially revealed, similar to Soo and Rottman's (2020) study, but with an interrupted time-series graph. We also tested 2 sequential trial-by-trial formats one in which the outcome was presented as a bar and another as a number, similar to Soo and Rottman's (2018) study.

### 1.6. *Timeframe of learning*

So far we have addressed 2 of our main questions: the accuracy of judgments about interrupted time series and the role of presentation format. Our third question is about how well people learn



and make judgments about interrupted time series data when presented in a format that mimics experiencing a sequence of events over a moderately long timeframe, like learning in one's everyday life.

We often make a change in our daily life to see if it produces a desired result (e.g., starting an antidepressant, or cutting out gluten from one's diet), which is essentially an informal interrupted time series situation. In such situations, people usually do not have access to all the data at once (e.g., a graph), but instead need to observe the events over a series of days. This may lead to a particular challenge for memory. Building off the example of starting an antidepressant, one would need to be able to compare how they felt during the period of time before versus after starting the antidepressant. If the antidepressant was started several days or weeks ago one may not have a very good memory for the period before the antidepressant.

Because existing theories of causal learning are primarily focused on situations in which the observations are randomized and do not form temporal trends (e.g., Cheng, 1997; Griffiths and Tenenbaum, 2005; Hattori and Oaksford, 2007), there is not much existing theory to explain how temporal trends are learned and remembered. Thus, in the introduction, we will speculate about potential cognitive processes and how they may be impacted by long timeframe learning, and in the general discussion we will discuss implications for theories of causal learning more broadly. At a minimum, what must be remembered to make judgments similar to ITSA is to be able to learn the trend line (level and slope), both before and after the intervention. When the observations are experienced in a short timeframe working memory may be sufficient to accurately accumulate estimates of these trends. However, learning these trends and remembering them when the experiences are distributed over 14 days may be challenging, especially remembering the level and slope of the pre-intervention period, which by the end of the study took place 7–14 days earlier.

One specific hypothesis is that when the learning data are experienced over many days, people may rely more on the post-intervention trend heuristic, because this heuristic can be used even if a learner has forgotten the pre-intervention events. For example, a patient may think a medication is very effective when they find that they take the medication every single day and their symptom also improves every single day, ignoring the fact that the symptom had already started to improve before they started the medication and ignoring the possibility that the symptom might continue to get better even without the medication. Other hypotheses are discussed in the introduction to Experiment 3.

Since it is unknown how well people can assess interrupted time series from their observations in real life, we conducted an 'ecological momentary experiment' that presented one experience to participants each day for 14 days. There have only been a few prior ecological momentary experiments. The first found that people were able to learn causal relationships as well when they learned one trial per day as when the data were presented back to back within a few minutes (Willett and Rottman, 2021). In a second study, we also found that delays between the cause and effect do not substantially affect causal learning; people were able to learn cause-effect relations with delays of up to 21 hours with minimal differences to no delay (Zhang and Rottman, 2021). However, people were not always rational in the long-timeframe causal learning. People inferred an illusory correlation when the cue and outcome often occurred together, even though there was no correlation between the cue and outcome, similar to in short timeframes (Willett and Rottman, 2021). In a third study we found that, when learning about 2 potential causes simultaneously, people had difficulty controlling for the stronger cue when assessing the influence of a weaker cue in the long timeframe, though they were not especially good in the short timeframe either (Willett and Rottman, 2020).

However, in the previous studies, the trials were presented in a randomized order so there were no temporal trends in the data. In contrast, the goal of the current research was to investigate if people could learn about temporal trends, particularly changes in level and slope, in a long timeframe context. Additionally, in the previous studies, the random trial order meant that participants could perform fairly well even if they only remembered fairly recent events, whereas in an interrupted time series design participants need to remember what happened before the intervention (7–14 days earlier) to accurately infer changes in the level or slope.

### 1.7. Current study

The main goal of the current research was to investigate how people infer causation from various interrupted time series datasets. We have 3 main questions: (1) Do people's causal judgments correspond to ITSA when making judgments about AB-interrupted time series data in the short timeframe? (2) Do different presentation formats moderate Question 1? (3) When people make judgments about AB interrupted time series data when presented over a long timeframe, mimicking real-life learning, do their judgments correspond to the inferences of ITSA?

In Experiment 1, we generated a spectrum of situations that either do or do not have pre-intervention slopes, and have various post-intervention changes in the level or in the slope to address Question 1. We also explored individual differences in people's causal judgments. Moreover, we examined the influence of 4 presentation formats, including 2 graph formats and 2 trial-by-trial formats.

In Experiment 2, we focused on the trial-by-trial paradigm and a subset of interrupted time series conditions. Importantly we used a between-subject design to eliminate any potential inference between time series conditions.

In Experiment 3, we investigated how people would learn cause-effect relations from interrupted time series in their daily life. We implemented the experiments in a long timeframe in which we presented one trial per day.

## 2. Experiment 1

The main goal of Experiment 1 is to address Question 1 about whether people's judgments about interrupted time series correspond to ITSA and Question 2 about whether they are moderated by different presentation formats. In Experiment 1, we included 9 interrupted time series situations that fit with the generative model from ITSA; they either do or do not have pre-intervention slopes, changes in the level, or changes in slope. In addition, we also presented the time series data in 4 different formats, comparing summarized graphs with sequentially and dynamically presenting formats.

### 2.1. Method

#### 2.1.1. Participants

A total of 402 participants were recruited on Mechanical Turk (168 females); the preregistration said we would recruit 400 and 2 additional participants did the study without submitting the HIT. All participants had an overall HIT Approval Rate greater than or equal to 95%. The experiment lasted 10–15 minutes and participants were paid \$2. We excluded 11 participants from data analysis because they reported they experienced a technical error during the experiment.

#### 2.1.2. Design and stimuli

The experiment was a 4 (presentation formats; between-subjects)  $\times$  9 (interrupted time series conditions; within-subjects) mixed design.

##### 2.1.2.1. Data for 9 time series conditions

Figure 2 presents the 9 time series datasets<sup>2</sup> that we investigated in the current experiment. Figure 2 also shows the predictions made by the various theories.

The 9 conditions can be divided into 3 groups. The *No Influence* group includes 2 conditions. Condition A, which we call the 'flat' condition, has zero pre-intervention slope, and no slope change nor a level change. All theories predict no causal relationships for this condition. Condition B is similar to White's (2015) condition in which there is a nonzero pre-intervention slope but no slope or level change. Only ITSA predicts there is no causal relationship in this condition.

<sup>2</sup>We did not investigate situations that have changes to both the level and the slope. In such situations, the change in slope and level could be in opposite directions (e.g., positive level change but negative slope change), in which case it is not clear how to answer the dependent measures about whether the intervention had a positive or negative causal influence.

The *Level Change* group includes 3 conditions in which there is a level change. Condition C has a flat pre-intervention trend. Condition D has a nonzero pre-intervention slope that is in the same direction as the level change (e.g., in Figure 2 a positive pre-intervention slope and a positive level change), which we call ‘congruent’. Condition E has a pre-intervention slope in the opposite direction as the level change (e.g., in Figure 2 a negative pre-intervention slope and a positive level change), which we call ‘incongruent’. Condition E is especially useful because the models make very different predictions.

The *Slope Change* group includes 4 conditions with a slope change. Condition F has a flat pre-intervention trend, Condition G has a pre-intervention slope that is congruent with the causal influence (e.g., in Figure 2, a positive pre-intervention slope and a positive change in slope). Condition H has an incongruent pre-intervention slope (e.g., positive pre-intervention slope but negative slope change). In Condition I, the post-intervention slope is 0. Condition I is useful because, similar to Condition E, the models make very different predictions. According to ITSA participants should infer positive causal efficacy. According to the After-minus-Before model, they should infer negative, and according to the post-intervention-trend model, they should not infer any influence of the cause.

For generality, we included 2 parallel datasets for each time series condition, which simply involved flipping the *Y*-axis. In the positive datasets, the intervention caused the outcome to get higher, whereas in the negative datasets, the intervention caused the outcome to get lower (for Conditions A there is no difference between positive and negative, and for Condition B the negative version was a negative slope). For data analysis, all judgments in the negative datasets were reverse coded and the 2 parallel conditions were analyzed together.

We also introduced a small amount of noise to the datasets to make them more realistic (which was not presented in Figure 2). We added noise to the baseline datasets and then rounded them to whole numbers. We created 20 predetermined noise sequences. Each of the 20 noise sequences used the following set of noise both for the pre and post-intervention phase:  $[-2, -1, -.5, 0, .5, 1, 2]$ . The noise was randomly ordered among those 7 trials, however, for all 20 sequences we verified that even after adding the noise ITSA produced the correct inferences. Specifically, for all the coefficients listed under the ‘generative model’ in Figure 2, if the coefficient was supposed to be positive or negative, they were, and the *p*-values were less than  $10^{-7}$ . For cases that were supposed to have coefficients near 0 all of the *p*-values were larger than 0.1.

#### 2.1.2.2. Cover story

The cover story read as follows:

“Please imagine that you work for a medicine company that is testing a whole variety of new medications on a wide variety of symptoms in order to test whether the medications improve or worsen the symptoms. Specifically, they conducted 9 studies to test the efficacy of 9 new medicines for 9 different kinds of symptoms.

In the current study, you will review data from these 9 experiments. In each experiment, you will view data for a single patient over 14 days. You will see whether or not the patient took the medicine each day and the level of their symptom. Each experiment has a different kind of medicine, each identified with a code (e.g., SNP27), and a different kind of symptom.

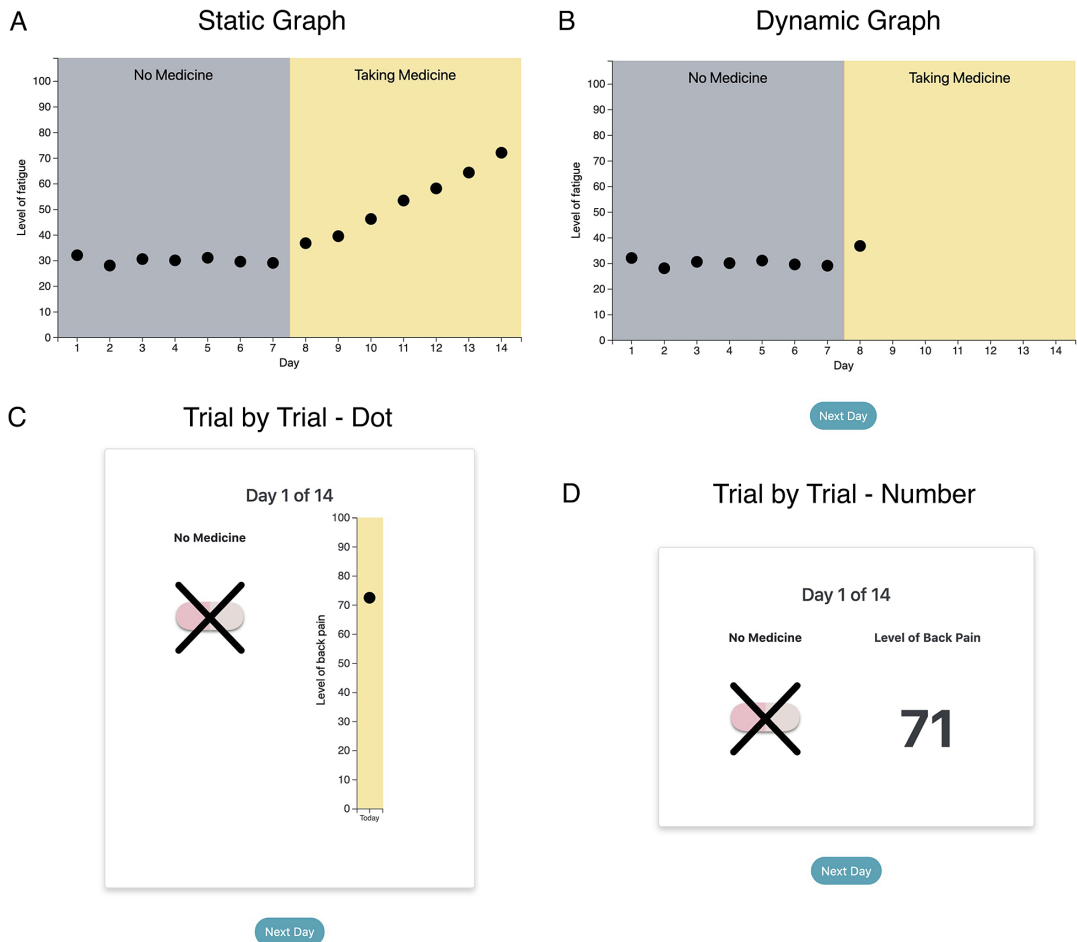
For the first 7 days, the patient did not receive any medicine. From Day 8 to Day 14, the patient received the medicine. The patient’s symptoms were recorded once per day on a scale from 0 to 100. 100 means that the patient had very bad symptoms and 0 means that the patient had no symptoms.

Please remember that the medications could make the symptoms better, or worse, or have no influence on the symptoms”.

#### 2.1.2.3. Four presentation formats

Figure 3 shows the 4 presentation formats (static graph, dynamic graph, trial-by-trial dot [hereafter TbT-dot], or trial-by-trial number [hereafter TbT-number]).

In the static graph condition, all 14 observations were presented in a dot chart. The pre- and post-intervention periods were indicated by different colored backgrounds. The dynamic graph condition



**Figure 3.** Four presentation formats in Experiment 1.

was identical to the static graph condition, except that a data point was added to the graph each time participants clicked a button.

In the TbT-dot condition, participants saw one observation per trial. Each trial contained an icon that indicated the status of the intervention (medicine) on the left, and a bar (a narrow dot chart) with a dot that showed the level of the outcome. Participants clicked a button to see the next trial. The TbT-number condition was identical to the TbT-dot condition, except that the dot chart was replaced by a number of the level of the outcome. To avoid participants from going through the observations too fast, there was at minimum a 1-second waiting time between each click in the dynamic graph and 2 trial-by-trial conditions.

In all formats but the static format, there was an alert between trial 7 and trial 8 to remind participants that the patient has started taking the medication. In the static graph format, the dependent measure questions were shown at the same time with the graph. In the dynamic graph format, the questions were shown below the graph once all 14 data points were revealed. In the 2 trial-by-trial formats, the questions were shown on a new page after the 14 trials.

### 2.1.3. Dependent measures

After participants reviewed the observations in a dataset, they answered 2 questions about the influence of the medicine.<sup>3</sup> First, we measured participants' views about the 'causal strength' of the cause by

<sup>3</sup>After the two questions, participants also were asked to explain how they answered the two questions, though this open-ended question is not analyzed in the current research.

asking ‘Did taking [medicine] cause the [symptom] to get better or worse?’ on a 9-point scale: 4 indicates the medicine caused the symptom to get much worse—higher, 0 indicates the medicine had no influence on the symptom, and –4 indicates the medicine caused the symptom to get much better—lower.

Second, we asked ‘Do you think this patient should continue to take the medicine to treat the symptom?’ on a 9-point scale: 4 indicates ‘should definitely stop taking the medicine’, 0 indicates ‘unsure whether to continue or to stop’, and –4 indicates ‘should definitely continue to take the medicine’. If a participant thinks that the medicine is helpful, presumably the patient should continue to take it. This ‘future use’ question was designed to be an alternative way to enquire about causal efficacy but without having to mention causality directly, which is linguistically complex (Wolff, 2007).

## 2.2. Results

The analysis follows our pre-registered plan (<https://osf.io/uzt37>). We collapsed<sup>4</sup> 2 parallel datasets of each time series condition for data analysis, which means that according to ITSA the judgments should be 0 for Conditions A and B, and should be positive for all the other conditions. The Summary of Results column in Figure 2 shows a simplified summary of the results, which can be compared against the model predictions. Figure 4 depicts all the results and Table B1 shows inferential statistics. We provide *p*-values, effect sizes, and Bayes Factors.

### 2.2.1. Causal strength and future use judgments

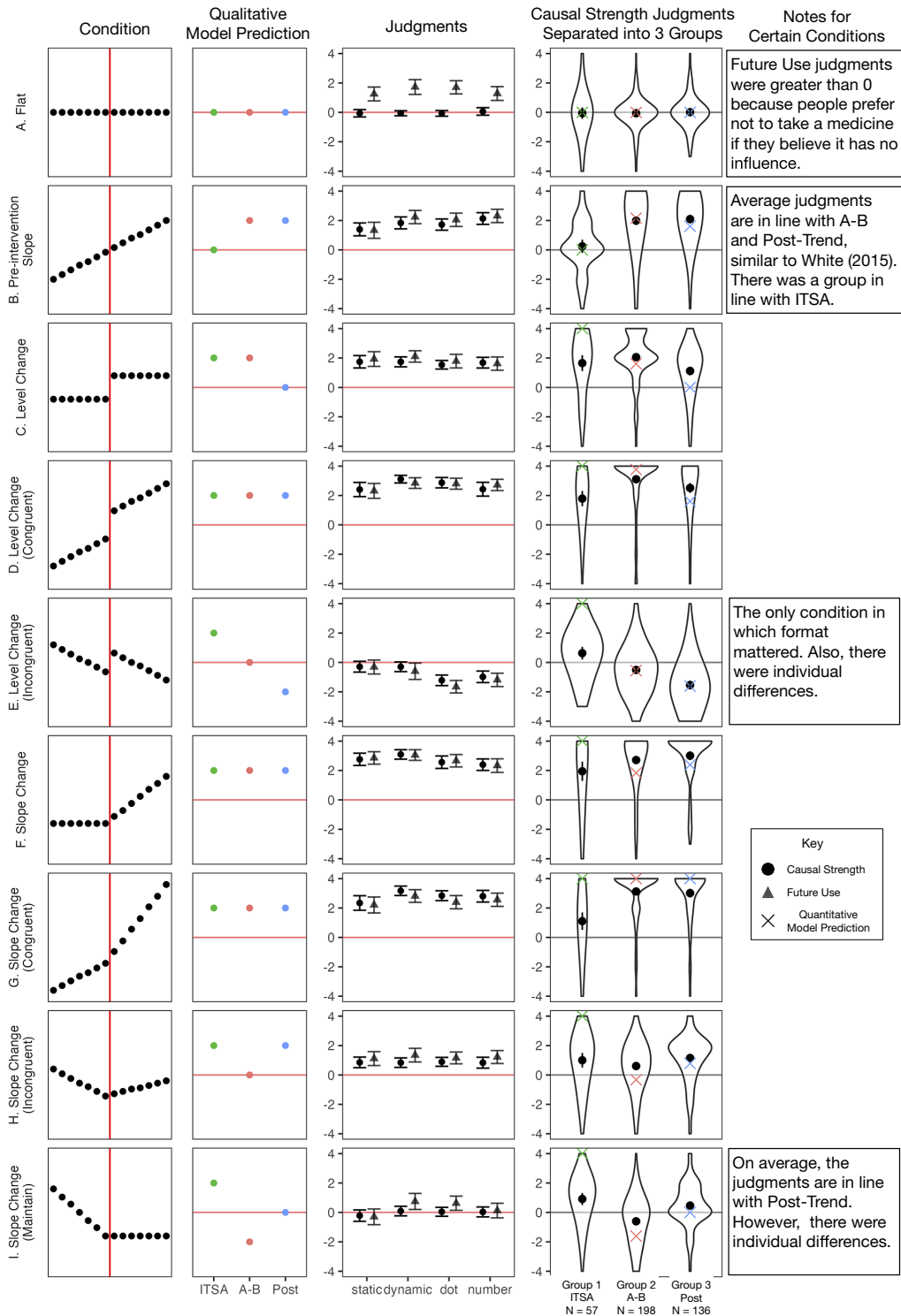
First, we compared the causal strength and future use judgments against 0 for each time series condition and format to see if the judgments fit the predictions of ITSA. The details of the statistical results are in Appendix B. The participants’ judgments were in line with ITSA for conditions A, C, D, F, G, and H. In Conditions C, D, F, G, and H, all judgments were appropriately above 0 ( $ps < .001$ , BFs > 100, Cohen’s *d* ranged from .72 to 2.41). In Condition A, participants appropriately gave causal strength judgments around 0 ( $ps > .05$ , BFs ranged from 1/3 to 1/10, Cohen’s *d* ranged from .05 to .07). However, the future use judgments were positive ( $ps < .001$ , BFs > 100, Cohen’s *d* ranged from .54 to .75), implying that they should stop taking the medicine. Though this differs numerically from causal strength, it makes sense if participants believe that if a medication has no benefit, it should not be used, which would lead to positive judgments instead of 0. Note that this comment about future use applies to other conditions as well; it is understandable that future use judgments may generally be more positive than causal strength.

Participants’ judgments differed from the predictions of ITSA in Conditions B, E, and I. In Condition B, the causal strength judgments were higher than 0 ( $ps < .001$ , BFs > 100, Cohen’s *d* ranged from .49 to 1.09), even though they should be 0 according to ITSA. This finding is somewhat similar to White’s (2015) finding that people infer that an intervention made a difference even when the change in the outcome started before the intervention (see an example in Figure 1C). This is consistent with the After-minus-Before model and the Post Trend model.

In Condition E, the judgments for the static and dynamic graph were close to 0 ( $ps > .05$  for all graph conditions but the future use judgments for the static graph condition;  $p = .035$  for the static graph condition; BFs < 1, Cohen’s *d* ranged from .13 to .22). The judgments for the trial-by-trial formats were below 0 ( $ps < .001$ , BFs > 100, Cohen’s *d* ranged from .51 to .77). Thus, these are most consistent with the After-minus-Before model and the Post Trend model but not ITSA.

In Condition I, the causal strength judgments were close to 0 ( $ps > .05$ , BFs ranged from .11 to .21, Cohen’s *d* ranged from .02 to .11), even though they should be positive according to ITSA. The future use measures were close to 0 for 2 out of the 4 conditions (Static:  $p = .265$ , BF = .21, Cohen’s *d* = .11;

<sup>4</sup>The comparison between positive and negative datasets is not of our interest for this study. There were some significant differences in the future use judgment between positive and negative datasets, which is likely due to a bias not to take a medicine. For example, in Condition A in which the medicine does nothing, most participants suggested stopping taking the medicine rather than the middle of the scale. However, this question is not of importance to the analysis, so we collapsed the positive and negative datasets.



**Figure 4.** Qualitative model predictions, means and 95% CIs for both dependent measures, and causal strength judgment were separated into 3 groups based on best fit to the 3 models in Experiment 1.

Note: The qualitative model predictions are the same in Figure 2. The quantitative model predictions were used for the individual difference analysis, indicated as crosshairs in the figure. 'A – B' means 'After-minus-Before' model, and 'Post' means 'Post Trend' model.



Number:  $p = .619$ ,  $BF = .13$ , Cohen's  $d = .05$ ), and were a bit positive for the other 2 but with weak BFs (Dynamic:  $p = .008$ ,  $BF = 3.5$ , Cohen's  $d = .27$ ; Bar:  $p = .012$ ,  $BF = 2.45$ , Cohen's  $d = .26$ ). Given that most of the judgments were near 0, they are most consistent with the Post Trend model.

The Summary of Results column in Figure 2 summarizes the findings qualitatively. None of the theories can explain all of the results. There are some conditions that are best explained by the After-minus-Before heuristic (graph formats in Condition E) and some that are best explained by the Post Trend heuristic (Condition I). As explained below, there are also some findings that are best explained by ITSA.

### 2.2.2. Effects of presentation format

We conducted ANOVAs to compare 4 presentation formats within a time series condition. Table B1 shows the statistical results in the 'ANOVA' columns. For 8 out of the 9 conditions, there were no reliable effects of presentation format; most of these had  $ps > .05$  and BFs  $< 1$ , while a few had slightly significant  $p$ -values but still weak BFs.

In contrast, there was a main effect of presentation format in Condition E for both causal strength ( $p < .001$ ,  $BF = 88.63$ ,  $R^2 = .05$ ) and future use ( $p < .001$ ,  $BF = 38.67$ ,  $R^2 = .05$ ). The judgments with the static and dynamic graph formats were close to 0 but the TbT-dot and TbT-number formats were less than 0. We tested all pairwise comparisons and report  $p$ -values with a Tukey correction and BFs. For causal strength, the static graph group was higher than the TbT-dot ( $p = .002$ ,  $BF = 53.2$ , Cohen's  $d = .5$ ) and TbT-number ( $p = .042$ ,  $BF = 3.8$ , Cohen's  $d = .36$ ). The dynamic graph group was also higher than the TbT-dot ( $p = .002$ ,  $BF = 95.5$ , Cohen's  $d = .53$ ) and TbT-number ( $p = .043$ ,  $BF = 4.1$ , Cohen's  $d = .38$ ). There were no differences for the remaining comparisons ( $ps > .05$ , BFs between .16 and .22, Cohen's  $d$  between .002 and .12). The future use judgments had a fairly similar pattern. The judgments with the static graph and dynamic graph were significantly higher than with the TbT-dot format (static vs. TbT-dot:  $p < .001$ ,  $BF > 100$ , Cohen's  $d = 0.59$ ; dynamic vs. TbT-dot:  $p = .011$ ,  $BF = 9.42$ , Cohen's  $d = 0.42$ ). There was no difference between the 2 graph formats or 2 trial-by-trial formats (all  $ps > .05$ , BF between .20 and .43, Cohen's  $d$  between .11 and .21). One difference compared to the causal strength judgments is that for future use there was no reliable difference between dynamic graph versus TbT-number ( $p = .323$ ,  $BF = .52$ , Cohen's  $d = .23$ ) or static graph versus TbT-number ( $p = .056$  but  $BF = 3.9$ , Cohen's  $d = .38$ ).

In summary, for most of the conditions there was no reliable influence of format. In Condition E, participants' judgments deviated from ITSA in all formats. They should have inferred a positive relation due to the positive level change (or a negative relation for the flipped version), but in the 2 graph conditions, they inferred no influence, which corresponds to the predictions of the After-minus-Before heuristic. Additionally, in the 2 trial-by-trial conditions they inferred a negative influence despite the positive level change (or flipped), which corresponds to the predictions of the Post Trend heuristic.

### 2.2.3. Individual differences in causal judgments

Here we report an analysis of individual differences; we thank anonymous reviewers and the editors for suggesting this. We first calculated the predictions for the 3 models, Post-Trend, After-minus-Before, and ITSA, for all 9 conditions. We coded the predictions for ITSA as 1 if there was a change in level or a change in slope (Conditions C–I), and 0 if there was neither (Conditions A and B). For the other 2 models, we used the actual values of the post-trend slope and the actual mean of the post-trend minus the mean of the pre-trend. (Note, these quantitative predictions look slightly different from the 'Qualitative Model Prediction' Column but were indicated in the 'Causal Judgments Separated into 3 Groups' Column in Figure 4. The predictions in 'Qualitative Model Prediction' Column in Figure 4 are just meant to capture qualitatively if the prediction is positive, negative, or 0. Here the predictions for Post-Trend and After-minus-Before are quantitative. For example, the post trend was the strongest for Condition G, then F, H, and finally I.)

We then ran regressions for each participant to predict their causal strength judgments for the 9 conditions. We ran 3 regressions, one for each model. (We only ran this analysis for the causal strength

judgments because, as already explained, it makes sense that the future use judgments can deviate from the predictions of these models if people tend to not want to use a medicine if it has no influence.) We then identified which model had the best fit for each participant; Because there is only one predictor in each of the 3 models, it does not matter which metric (e.g.,  $R^2$ , AIC, BIC) is used.

We found that 198 participants were best fit by After-minus-Before, 136 by Post-Trend, and 57 by ITSA. The ‘Causal Judgments Separated into 3 Groups’ Column in Figure 4 visualizes the judgments for the participants in each group. For the most part, the judgments within each group do seem to fit with the predictions of the model with a few exceptions. First, in Group 3 (Post Trend), the judgments for Condition C were positive even though according to the Post-Trend model they should be 0; participants were probably responding to the very obvious change in intercept for Condition C. In Group 2 (After-minus-Before), the average judgments for Condition H were slightly positive, but this model predicts them to be close to 0 (technically slightly negative). In Group 1 (ITSA), the judgments line up fairly well with ITSA qualitatively.

In sum, it appears that some participants’ judgments are better explained by each of the 3 models, and mostly by After-minus-Before and Post-Trend. However, we want to be clear that we are not proposing that participants in these groups were literally using the respective model just that it is the best-fitting model out of the 3. Furthermore, this approach revealed that there is variance in the judgments for participants within a group, and a few of the averages deviated from the predictions of the model, which means that it is still likely that they are using some mixture or being somewhat inconsistent, or using other approaches aside from these 3.

We also did another sort of individual difference analysis using latent profile analysis (LPA). The key difference is that approach is more bottom-up, and does not assume the 3 models as priori. That LPA approach was in line with what we had preregistered, though in the end, we are presenting this approach since it more directly connects the results with the models; the results from the LPA approach are available on OSF. In the end, the conclusions of the 2 approaches are fairly similar, that it appears that there are some participants who are using strategies fairly close to these 3 models, but also that it is likely that there are some mixtures of strategies or other strategies being used.

### 2.3. Discussion

This is the first experiment to systematically investigate causal learning from interrupted time series data that exhibits a change in slope or level. Figure 2 provides a summary of the results compared to the models. Participants accurately judged time series situations A, C, D, F, and H. In contrast, they made judgments that disagreed with ITSA for situations B, E, and I. There is no simple rule that delineates the situations in which people’s judgments tend to agree or disagree with ITSA.

When judged in the aggregate, in all 9 situations either the After-minus-Before or the Post Trend model predicts participants’ average judgment, and in some cases, only 1 of these 2 is consistent with the judgments; there are no situations in which only the ITSA model predicts participants’ judgments. Conditions E and I are interesting because some of the judgments are only consistent with the Post Trend model, but at the same time Condition C and some of the judgments in Condition E are inconsistent with the Post Trend model. It is possible that for the situations that participants’ judgments aligned with ITSA, they could have been using approaches similar to After-minus-Before or Post Trend. Specifically, in conditions A, D, F, and G, participants’ judgments were consistent with all 3 models.

We separated participants’ causal strength judgments into 3 groups based on the best-fit model to examine if there were groups of participants who responded in different ways. This analysis found that there was a subset of about 14.5% of participants who make judgments akin to ITSA in all conditions. There was also a group of participants who made judgments more in line with Post Trend (34.7%) and another who made judgments more in line with After-minus-Before (50.6%); however, for these, the latter 2 groups neither theory could explain all of the results.

With regards to format, the only reliable influence of format appeared in Condition E. In the static and dynamic graph condition for E, one possibility is that since the participants could see all the data at

once, they could more easily implement the After-minus-Before heuristic, which in this case revealed only a slight negative influence—close to 0. Another possibility in this condition is that participants were aware of both the positive level change (ITSA) and also the negative post-intervention slope (or flipped) and that they gave judgments near 0 because of the opposing views. In the trial-by-trial formats, participants tended to give negative judgments. One hypothesis is that they primarily focused on just what was happening during the intervention, not comparing the period during the intervention to the period prior to the intervention. It may be somewhat hard to remember the earlier period in these trial-by-trial conditions. Alternatively, it is possible that the trial-by-trial format encourages noticing the change from one trial to the next, which was negative.

### 3. Experiment 2

The main goal for Experiment 2 was to refine the methodologies from Experiment 1 before using them for Experiment 3, which tested how well people make judgments about interrupted time series data in a long timeframe situation. There were multiple methodological issues we wanted to address before doing the much more complicated long timeframe experiment. Most importantly, in Experiment 1 the 9 conditions were within subjects, so it is possible that participants learned about these interrupted time series situations over the course of answering questions about multiple cases and responded to them in a different way than they would if they were only exposed to one. To be more certain about the results in the short timeframe before running the long timeframe study, we ran Experiment 2 between subjects. Other smaller changes are explained in the methods.

#### 3.1. Methods

##### 3.1.1. Participants

We recruited 302 participants from Mturk with the same requirements as in Experiment 1. We dropped 19 participants who reported that they wrote down data to help their memory in the post-experiment survey. The data analysis included 283 participants (97 females).

##### 3.1.2. Design and stimuli

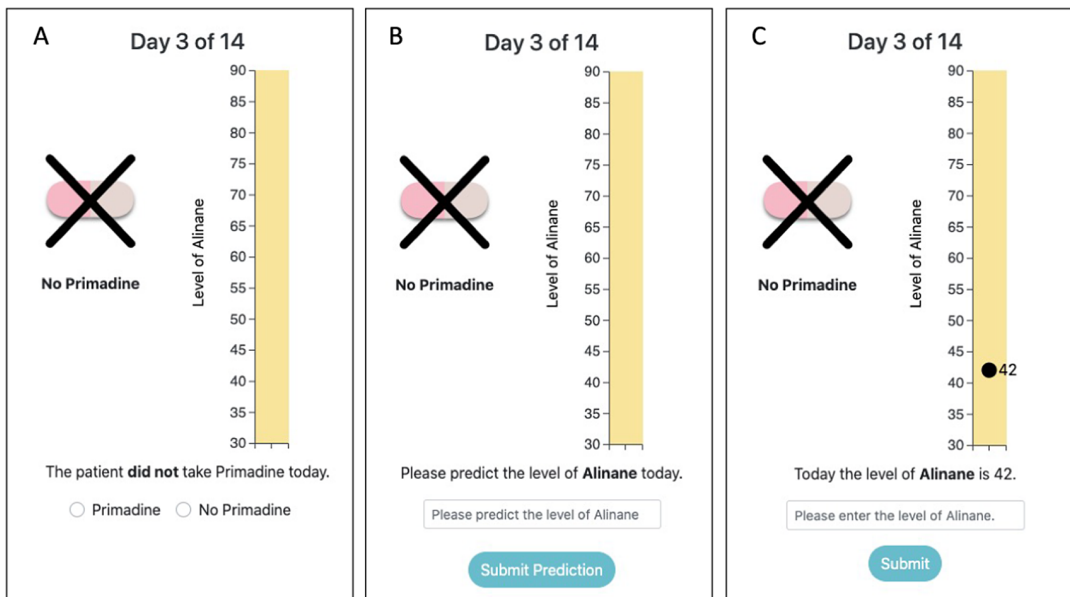
There were several changes in the design and stimuli of Experiment 2 from the previous experiment. First, Experiment 2 used a between-subject design in which each participant reviewed only one of the interrupted time series conditions.

Second, in Experiment 1 all 3 models agree for Conditions A, D, and G, and also agree with participants' judgments. We dropped these conditions because they do not add much to the results, and kept the remaining 6.

Third, for the remaining 6 conditions from Experiment 1, we used slightly different datasets. In the previous study, we told participants that the level of the outcome was in a range of 1–100. For any bounded variable, it is reasonable to assume that a cause-effect relation might be nonlinear (e.g., the cause might have a strong influence near the middle of the scale but barely any influence near an end). In the current study, we told participants that the level of the outcome could go from a minimum of 0 to a maximum of many hundreds, and the scale they saw only presented data between 30 and 90 to avoid getting close to the minimum of 0 to reduce concern about potential nonlinearities.

For Conditions E and G in Experiment 1, the After-minus-Before model was not exactly 0; it was slightly off from 0 (Even though for Experiment 1 the After-minus-Before model was not exactly 0, because it was so close we treated it as 0. The stimuli are available on OSF). In the current experiment, in order to compare the 3 models more precisely, we redesigned the datasets, and the After-minus-Before model gave a prediction of exactly 0 for both Conditions E and G. We also changed the noise sequence to  $[-2, -1, 0, 0, 0, 1, 2]$  because we used a baseline with whole numbers.

Fourth, we changed the cover story so that the effect was described as a chemical in the patient's blood instead of a symptom. In Experiment 1, we used a medicine-symptom cover story in which



**Figure 5.** An example trial in Experiments 2 and 3.

we told participants higher level of symptom means worse and lower means better, but this could be confusing as often higher is intuitively viewed as better. In order to reduce this potential confusion, in Experiment 2 we did not tell participants whether a higher level of the chemical was better or worse.

Fifth, because in Experiment 1 we found only a minimal influence of format, in Experiment 2 we only used one format. Because the goal of Experiment 2 was to further develop the methods we would eventually use in the long timeframe in Experiment 3, we used a trial-by-trial format; the goal of the long timeframe study is to investigate how people learn from a series of experiences spaced out over time.<sup>5</sup> Instead of using either the TbT Dot or Number formats, we combined the two formats (Figure 3).<sup>6</sup>

### 3.1.3. Within trial procedure

Figure 5 shows the procedure within each trial. First participants saw an icon and text that indicated the presence or absence of the cause, the medicine Primadine (Figure 5A). Participants verified if the patient took the medicine or not by clicking a button. Only after they responded correctly could they proceed. Second, they were asked to predict the level of the outcome, the chemical Alinane, by typing in a number between 30 and 90 (Figure 5B). Third, participants were shown the actual level of Alinane with both a dot and a number (Figure 5C). In order to ensure that they attended to the outcome, the participants had to verify the level of the chemical by typing the number before proceeding.

### 3.1.4. Dependent measures

We made a few changes in the dependent measures. We did not use the future use question from Experiment 1. Participants' judgments about future use largely agreed with their causal strength

<sup>5</sup>There are some circumstances in which a learner may see a graph each day, and the most recent day gets added into the graph, similar to the dynamic graph format from Experiment 1 (e.g., looking at graphs of the stock market in the newspaper each day). However, the main goal of Experiment 3 was to test learning in situations in which the learner must keep track of the experiences in memory, so we only tested the trial-by-trial condition.

<sup>6</sup>The reason for combining the two was largely for Experiment 3. Soo and Rottman (2018) found that visual stimuli were better than numbers for causal inference in a short timeframe. However, when data were spread over a longer period of time, it might actually be very hard to compare the levels of dots visually from one day to the next, or remember the numbers from day to day. We decided to include both formats to maximize the possibility of learning.

judgments, so they did not add much value. Additionally, because of participants' reasonable bias to suggest not using a medicine that has no influence, this measure is harder to compare to the predictions of the heuristics, for which no influence corresponds to 0 instead of a negative value.

We added a measure during the learning phase; we collected participants' trial-by-trial prediction of the outcome prior to seeing the true level of the outcome (Figure 5). This was not included in Experiment 1 because it could not be used for the static graph condition and we wanted all 4 formats to be comparable, however, we have used trial-by-trial predictions in other experiments and so have others (Well et al., 1988; Willett and Rottman, 2021).

After reviewing 14 trials, participants answered 3 questions. First, the *causal strength* question was modified slightly because the outcome was no longer described as better or worse. The new question asked, 'Did taking the medicine Primadine cause the level of Alinane to get higher or lower?'. Participants rated the causal strength on a 9-point scale with the following labels: 4 'the medicine caused the level of Alinane to get much higher'; 3 '... somewhat higher'; 0 'Primadine has no influence'; -2 '... somewhat lower', and -4 '... much lower'. On the next screen, they explained how they answered the previous question but we did not analyze participants' explanations in this article.<sup>7</sup>

Second, we included a new *predictive strength* measure, which includes 2 parallel questions in which we asked 'Imagine that "tomorrow" (Day 15), the patient [continues to take/ stops taking] Primadine. What do you think the level of Alinane will be?' Participants typed in a number between 0 and 200. We took the difference in their answers to the 2 questions as the predictive strength; if participants believe that the medicine has a strong influence then there should be a big difference.

Third, we added a new measure of participants' memories for the levels of chemical Alinane (outcome) over the 14 trials of the study. Participants were shown a blank graph of the 14 trials. For each of the 14 trials, participants recalled the level of the outcome and typed in a number between 30 and 90. As they entered numbers for each trial a dot appeared on the graph for each day. After they recalled the level of the outcome for all 14 trials, a 'Submit' button appeared allowing them to proceed. For clarity, during the learning phase, participants only saw the outcome for one trial at a time, but during the memory test, they created a graph of their memories of all 14 trials.

## 3.2. Results

The analysis follows our pre-registered plan (<https://osf.io/ue37t>) except where mentioned.

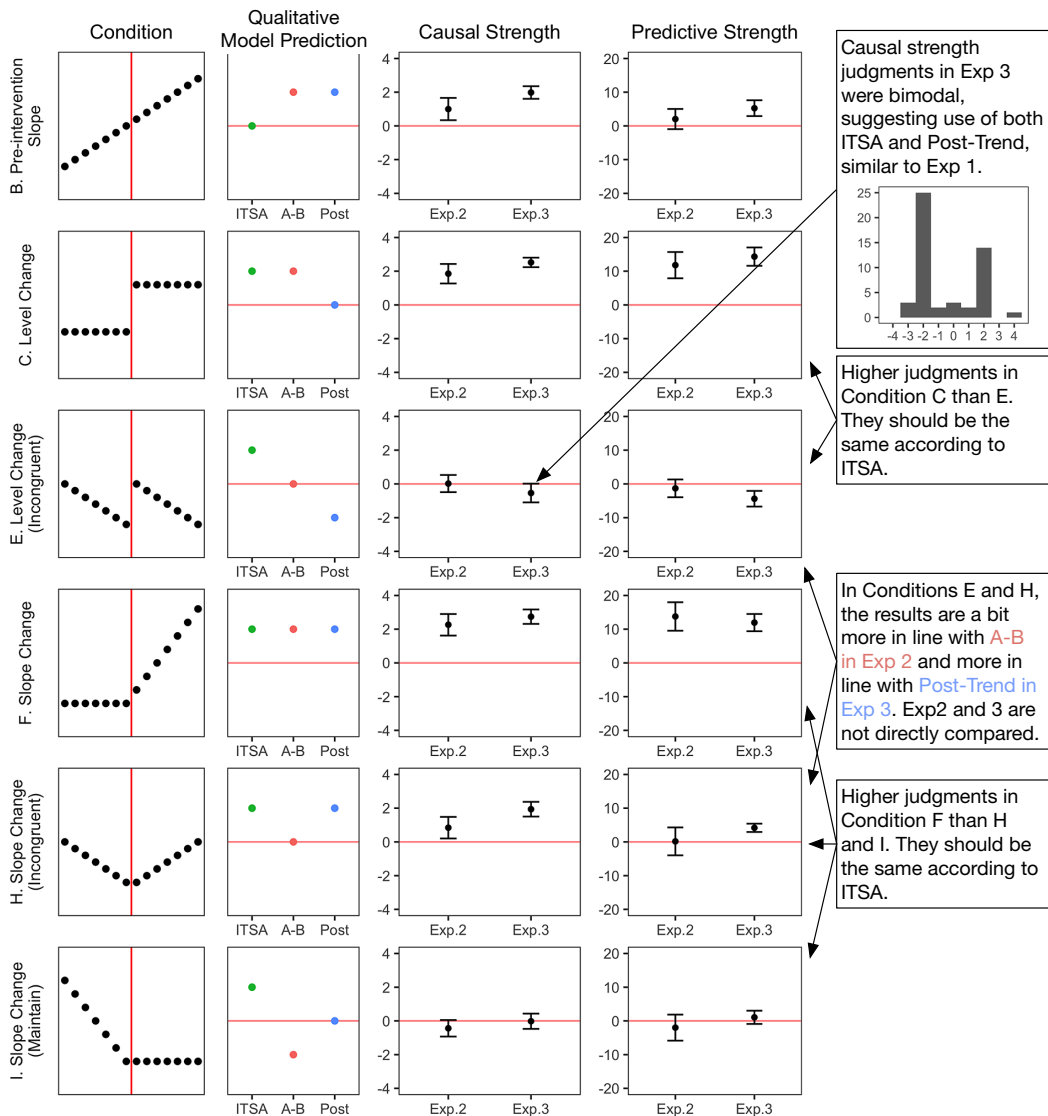
### 3.2.1. Causal and predictive strength judgments

The causal strength and predictive strength judgments are shown in Figure 6. We first compared conditions within the level change group and the slope change group. According to ITSA, the level change and slope change regression coefficients are identical for the 2-level change conditions and for the 3 slope change conditions. Thus, according to ITSA the causal strength and predictive strength judgments should be the same across the conditions within the same group. The inferential statistics for these comparisons are shown in the bottom half of Table 1. Participants gave higher judgments for C than E, and higher judgments for F than H and I. This confirms that conditions that should be the same according to ITSA have different judgments.

We then compared the causal strength and predictive strength judgments against 0 for each condition using *t*-tests (top half of Table 1). The judgments in Conditions C, F, and I were in line with Experiment 1.

Conditions B, E, and H revealed some similarities and some differences to Experiment 1. In Condition B, the predictive strength judgments were not significantly different than 0, which is predicted by ITSA. However, it is possible that with more data it would be significantly positive

<sup>7</sup>Most of the explanations were not very clear. A few participants described the trend but did not explain their reasoning. Some gave reasoning that deviated from ITSA and some gave reasoning akin to ITSA. Coding these responses was only listed as an exploratory analysis in the registration.



**Figure 6.** Means and 95% CIs for judgments in Experiments 2 and 3.

*Note:* ‘A–B’ means ‘After-minus-Before’ model, and ‘Post’ means ‘Post Trend’ model. The histograms and figures with individual observations for each measure are available in OSF.

(Cohen’s  $d = .20$ ). In contrast, the causal strength judgment was greater than 0, similar to Experiment 1 and consistent with the After-minus-Before and the Post Trend models.

In Condition E, both judgments were close to 0 and in line with the After-minus-Before heuristic. This is inconsistent with the findings in Experiment 1 for which the causal judgments in the 2 trial-by-trial formats were below 0 and in line with the Post Trend heuristic.

In Condition H, the causal strength judgments were significantly higher than 0, which was in line with ITSA and post-intervention slope and Experiment 1. However, the predictive strength was close to 0, which can only be explained by After-minus-Before.

In sum, similar to what we found in Experiment 1, none of the theories can explain all of the results.



**Table 1.** Comparing judgments to 0, and pairs of conditions, in Experiments 2 and 3.

| Conditions                    | Experiment 2    |       |          |                     |      |          | Experiment 3    |      |          |                     |      |          |
|-------------------------------|-----------------|-------|----------|---------------------|------|----------|-----------------|------|----------|---------------------|------|----------|
|                               | Causal strength |       |          | Predictive strength |      |          | Causal strength |      |          | Predictive strength |      |          |
|                               | <i>p</i>        | BF    | <i>d</i> | <i>p</i>            | BF   | <i>d</i> | <i>p</i>        | BF   | <i>d</i> | <i>p</i>            | BF   | <i>d</i> |
| Comparing means against 0     |                 |       |          |                     |      |          |                 |      |          |                     |      |          |
| B                             | .004            | 8.98  | .45      | .177                | .38  | .20      | <.001           | >100 | 1.5      | <.001               | >100 | .64      |
| C                             | <.001           | >100  | .94      | <.001               | >100 | .89      | <.001           | >100 | 2.53     | <.001               | >100 | 1.49     |
| E                             | .935            | .16   | .01      | .322                | .25  | .14      | .057            | .88  | .28      | <.001               | 64.8 | .54      |
| F                             | <.001           | >100  | 1.05     | <.001               | >100 | .97      | <.001           | >100 | 1.82     | <.001               | >100 | 1.33     |
| H                             | .011            | 3.62  | .4       | .94                 | .16  | .01      | <.001           | >100 | 1.27     | <.001               | >100 | .96      |
| I                             | .078            | .69   | .25      | .304                | .26  | .15      | .93             | .15  | .01      | .285                | .27  | .15      |
| Comparing pairs of conditions |                 |       |          |                     |      |          |                 |      |          |                     |      |          |
| C versus E                    | <.001           | >100  | .98      | <.001               | >100 | 1.15     | <.001           | >100 | 1.97     | <.001               | >100 | 2.10     |
| F versus H                    | .002            | 15.56 | .66      | <.001               | >100 | 0.97     | .010            | 4.31 | .53      | <.001               | >100 | 1.10     |
| F versus I                    | <.001           | >100  | 1.39     | <.001               | >100 | 1.13     | <.001           | >100 | 1.78     | <.001               | >100 | 1.35     |

### 3.2.2. Trial-by-trial predictions and memory

Figure 7 shows the means of the trial-by-trial predictions and the memories of the 14 trials. For both measures, we collapsed the positive and negative datasets by flipping the negative condition over the median of the true levels in each time series condition.

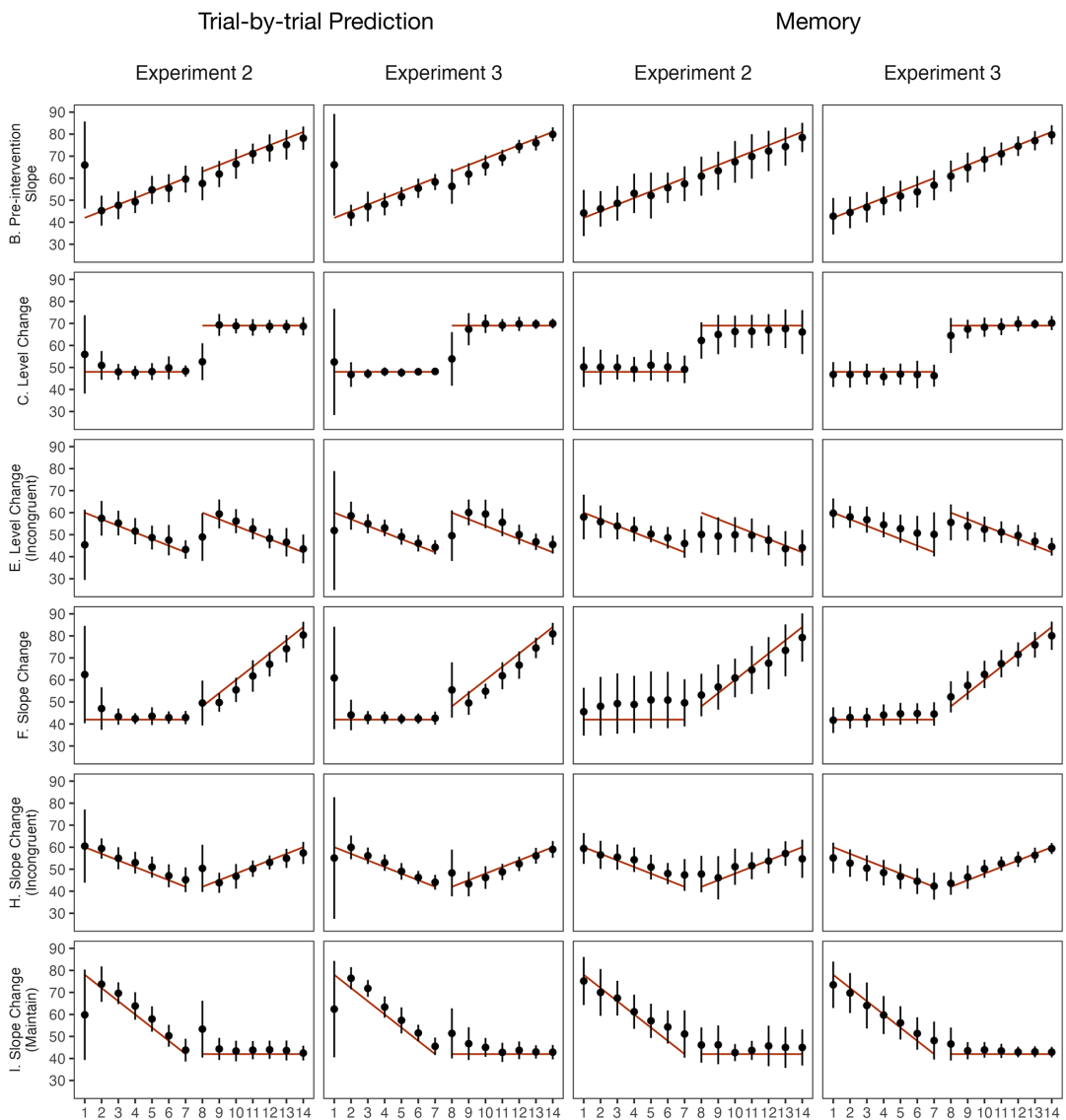
For each trial, participants predicted the outcome before they saw the actual level of the outcome. The predictions reflect participants' ability to remember prior trials and use them for predicting future trials. As can be seen in Figure 7, in all conditions participants were fairly accurate at making predictions both during the initial trend prior to the intervention and after the intervention. For example, even for Trials 2 and 3 during the initial period, and Trials 9 and 10, right after the intervention started, their predictions were fairly good. Note that in Figure 7, the predictions for Trials 1 and 8 should not be focused on because participants have little to base these predictions on.<sup>8</sup> At minimum, the accuracy of the predictions reflects good *local* learning (e.g., using the last trial or the last few trials to predict the next one).

In contrast, the memory measure assessed whether or not participants had accurate memories of all 14 trials at the end of the study (see the general discussion for more discussion of this measure). On average, the memories were very accurate. In some of the conditions (e.g., E) the slope or level change in the participants' memories is not quite as strong as the true values, but these deviations were small.

We ran ITSA on participants' memories with a by-subject random intercept and by-subject random slopes for the time, level change, and slope change to verify that their memories fit with the true experiences.<sup>9</sup> The left side of Table 2 shows the ITSA results in Experiment 2. Out of the 18 regression coefficients for Experiment 2 in Table 2, all but one aligned with the true ITSA Coefficients in Figure 2.

<sup>8</sup>In Trial 1, participants have no prior evidence so they have nothing to base predictions on; they tended to be in the middle of the scale. Trial 8 was the first trial of the intervention, so participants have little to base their predictions except from a guess about whether the intervention would make a difference or not.

<sup>9</sup>In the pre-registration, we planned to run interrupted time series analyses on each participants' memories to see if there were changes in the slope or level for the memories recalled by each participant, and if these participant-level parameters predicted their causal judgments. However, we did not find reliable results that we could interpret in meaningful ways, so we put these results on OSF but do not discuss them in this article. The analysis in this paragraph just examined participants' episodic memories alone and was not preregistered.



**Figure 7.** Participants' average trial-by-trial predictions and memories of the outcome in Experiments 2 and 3. The error bars are standard deviations. The red lines indicate the true values of the stimuli.

*Note:* For Trials 1 and 8, participants have little to base predictions on, so it makes sense that these predictions are not very accurate.

Specifically, when the coefficients were supposed to be positive, they were significantly positive, and when they were supposed to be 0 they were not significant. The one exception was that in Condition I, according to ITSA, participants incorrectly recalled a significant level change when in fact there was no level change. That said participants' memories in Condition I in Figure 7 looks fairly accurate. The condition that looks the most inaccurate in Figure 7 is E. Though the ITSA results suggest that people correctly inferred a negative pre-intervention slope, a positive level change, and no slope change, their memories for Trials 8 and 9 were not especially accurate and suggests that perhaps they did not robustly understand that there was a positive level change. This finding fits with the fact that participants' causal judgments and prediction strength judgments were near zero.

**Table 2.** ITSA results on memories of the outcome in Experiments 2 and 3.

| Conditions                    | Experiment 2  |               |               | Experiment 3  |               |               |
|-------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                               | Pre. slope    | Level change  | Slope change  | Pre. slope    | Level change  | Slope change  |
|                               | ( $\beta_1$ ) | ( $\beta_2$ ) | ( $\beta_3$ ) | ( $\beta_1$ ) | ( $\beta_2$ ) | ( $\beta_3$ ) |
| B. Pre-intervention slope     | 1.87***       | -0.26         | 0.62          | 1.97***       | 2.31**        | 1.06**        |
| C. Level change               | -0.32         | 11.97***      | 0.61          | 0.29          | 15.28***      | 0.71**        |
| E. Level change (incongruent) | -1.50***      | 6.36***       | 0.04          | -1.59***      | 7.77***       | -0.05         |
| F. Slope change               | -0.26         | 0.87          | 3.51***       | 0.51*         | 3.71***       | 3.62***       |
| H. Slope change (incongruent) | -1.92***      | -1.02         | 3.11***       | -1.92***      | -0.11         | 4.34***       |
| I. Slope change (maintain)    | -3.08***      | -4.70***      | 3.05***       | -3.87***      | -1.60         | 3.26***       |

\* $p < .05$ .\*\* $p < .01$ .\*\*\* $p < .001$ .

In sum, these results suggest that participants had fairly good local learning and remembered the observations fairly accurately with a few exceptions.

### 3.3. Discussion

Experiment 2 investigated causal learning with interrupted time series data in a between-subject design and a trial-by-trial format. Though there were some minor differences, the results are fairly similar to Experiment 1, suggesting that the deviations from ITSA in Experiment 1 were probably not due to the within-subjects design. We found that none of the 3 models could explain all results but all models could predict some time series conditions.

Two important additions in Experiment 2 compared to Experiment 1 were the trial-by-trial predictions and the memory judgments. The trial-by-trial predictions suggested that participants were able to fairly quickly learn the pre-intervention trend and the post-intervention trend. The trial memories suggest that by the end of the 14 trials, participants had fairly accurate memories of the overall pattern. The fact that participants were fairly accurate in these trial-by-trial measures deviates from the causal strength and predictive strength measures, which differ from ITSA in multiple conditions. The most plausible explanation, therefore, is that participants' learning of the pattern of events is fairly accurate, but their overall assessment of what the pattern means for making a causal judgment does not comport with ITSA.

## 4. Experiment 3

The goal of Experiment 3 was to test the extent to which people judge interrupted time series in line with ITSA when experienced in a more realistic timeframe in which the observations were spaced out once per day for 14 days. In the introduction, we discussed some hypotheses for how learning might occur when the events are spread out over time, which could tax memory resources and lead to using more economical learning processes. Here we elaborate on these ideas.

One possibility is that when the learning data are experienced over many days, people may forget the earlier pre-intervention events, which could have 2 implications. First, if people have difficulty

remembering the earlier part of the sequence (e.g., Days 1–7), this could be revealed in their memories; participants may not remember if the pre-intervention trend was increasing, decreasing, or flat, and their initial memories may or may not cohere with their causal judgments. For example, in Condition I in Experiment 2, participants gave causal judgments near zero despite remembering the initial decreasing slope followed by a flat slope. Even if participants still give judgments near zero in the current study, they might forget about the initial decreasing slope in Condition I, or other conditions. Relatedly, people might rely more on the Post Trend heuristic for judging the efficacy of the intervention—just noticing whether the outcome is increasing or decreasing near the end of the 14 days.

Another possibility is that people could institute a learning process similar to RW. Associative/reinforcement learning is relied upon when working memory is overloaded in short timeframe situations (Collins, 2018). And because it is a robust and efficient learning system—it stores a small number of parameter estimates that get updated after each new experience—it is a good candidate for situations with high demands on memory and attention. If participants use a process like RW they would only need to remember 2 things—they would need to sequentially update associative weights for the background and the cause, and they would not need to remember the prior trials.

The predictions of RW are quite similar to those of the After-minus-Before model with a few exceptions (Figure 2). Given that the findings in Experiment 2 were already best explained by the After-minus-Before model, then this prediction would mean fairly little change from Experiment 1, with the exception that presumably in Condition I the judgments would be negative (which is what is predicted by both After-minus-Before as well as RW). In sum, this experiment cannot really differentiate between After-minus-Before and RW, but the judgments could become more consistent with both. If participants use an associative learning process like RW it is also possible, that their causal inferences do not change much but that their episodic memories are worse, as RW does not rely on episodic memory.

Lastly, another hypothesis is that the judgments in this long timeframe study may be similar to those in Experiments 1 and 2. Some of our prior studies that specifically tested for differences in causal learning between short and long timeframes have not found very robust differences if any (Willett and Rottman, 2020, 2021). And Wimmer et al.'s (2018) study found differences between spaced versus massed reinforcement learning after a delay but not immediately after learning. That said, something unique about this study is that the cause was absent for the first week, and then present for the second week, as opposed to intermixed like in the prior studies. This places a very different demand on memory; effective learning cannot just involve memories of the last few trials.

## **4.1. Methods**

### **4.1.1. Participants**

A total of 307 participants completed the study, and 300 participants (212 females) were included in the data analysis; 2 participants were excluded for reporting that they wrote down data and 5 were excluded due to a programming error. Participants were recruited through the Introduction to Psychology subject pool and the Pitt+Me platform from the University of Pittsburgh and paid advertising online. Participants had to be between 18 and 30 years old, own a smartphone and use it frequently, and stay in the US Eastern time zone during the experiment since the timing of the study was tied to the time zone. 103 participants came from the Introduction to Psychology subject pool and they received one hour of credit and \$10 for completing the study. The rest of the participants came from the other recruiting mechanisms and received \$20 for completing the study.

### **4.1.2. Procedure**

Experiment 3 used the same design and stimuli as Experiment 2 but the trials were spaced out once per day for 14 days. The entire study was conducted on participants' smartphones. The study contained an introduction session which happened over Zoom on Day 1, a 14-day learning task, and a final judgment task on Day 15. On Day 1, participants were introduced to the study by a research assistant over Zoom. After the introduction, participants read the cover story (same as Experiment 2) and completed the

first trial of the 14-day learning task. On the following 13 days, participants completed one trial per day. Each trial was presented the exact same way as in Experiment 2 (Figure 5). Participants received text messages at 10 AM, 3 PM, and 8 PM each day to remind them to complete the task. Once they completed the trial on a given day they no longer received the reminder messages until the next day. If a participant did not do a task on a given day, the task would be pushed back for a day. If one missed more than 4 days, the task would be terminated, and the data would be dropped. In the final sample, 16 of 300 participants missed 1 day and 1 participant missed 2 days, so the experimental protocol was completed with high fidelity.

On Day 15, participants did the final judgment task; they were asked to do the task in a quiet room (e.g., library or home) without interruption. Participants answered the same 3 questions as Experiment 2 on their phones. Before they exited the website, they completed a quick questionnaire in which we asked whether they wrote down any of the learning data and whether they noticed any bugs in the app. The entire study was run with a custom-designed website created with our PsychCloud.org framework.

## 4.2. Results

The analysis follows our pre-registered plan (<https://osf.io/36mwq>) except where mentioned.

### 4.2.1. Causal and predictive strength judgments

The causal strength and predictive strength judgments are shown in Figure 6. When comparing conditions within the level change group and the slope change group, we again found that C was higher than E, and F was higher than H and I (Table 1). This means that there are significant differences between conditions that ITSA predicts to be exactly the same.

The inferential statistics comparing judgments to zero are in Table 1. Participants' causal strength and predictive strength judgments were significantly higher than 0 and in line with ITSA, in Conditions C, F, and H. These findings are almost entirely in line with the prior studies except in Experiment 2 the predictive strength judgment was not above 0 for Condition H.

The judgments in Conditions B, E, and I were not predicted by ITSA. Like most of the prior experiments, the judgments were higher than 0 in Condition B, even though according to ITSA they should be 0. In Condition I, participants inferred close-to-zero judgments, which fit with the prior studies and can only be explained by the post-intervention trend heuristic.

In Condition E, Participants' causal strength judgments were slightly (not significantly) below 0 on average. However, upon looking at the histogram, it appeared that there was a bimodal distribution; 25 of 50 participants selected the  $-2$  option, in line with the post-intervention trend model, and 14 of 50 participants selected the  $+2$  option, in line with ITSA. This is in line with the individual difference analysis in Experiment 1. In contrast, participants' predictive strength judgments in Condition E did not show such bimodality and most participants gave negative predictive strength judgments, which fits with the post-intervention trend.<sup>10</sup>

### 4.2.2. Trial-by-trial predictions and memory

The descriptive data of the trial-by-trial predictions and memory judgments are shown in Figure 7. On average, participants' predictions were close to the true levels. This means that the participants were able to remember and utilize the outcome from at least the previous few days to predict the subsequent day (which we call local learning).

Figure 7 shows that the means of the memories of the outcome were also close to the true levels. Table 2 provides the ITSA statistics computed from participants' memories. In Condition E, H, and I, participants' memories were very accurate (though in Figure 7, the level change in Condition E was

<sup>10</sup>Perhaps another possibility is that participants saw that the post-trend repeated the pre-trend, and hypothesized that if the medicine was stopped a third repetition would occur. This reasoning would result in a negative judgment, though it is hard to justify why this repetition would happen from a causal perspective.

not as strong as the true values); participants appropriately recalled the presence or absence of a pre-intervention slope, level change, and slope change.

In the rest of the conditions, the memories were less accurate. In Condition B, participants incorrectly recalled slope and level changes, which correspond to the above-zero causal judgments. In Condition C, an incorrect slope change was recalled. In Condition F, participants incorrectly recalled a pre-intervention slope and a level change. That said, when looking at the averages in [Figure 7](#), the memories actually look fairly accurate in these conditions and the deviations appear to be minor.

### 4.3. Discussion

In Experiment 3, we tested causal learning with interrupted time series data in a more realistic setting. Similar to the first 2 short timeframe experiments, we found all 3 models could explain judgments in some conditions but not all.

[Figure 2](#) shows a summary of the findings for Experiment 3, which can be compared against the model predictions. Compared to Experiment 2, ITSA predicted the results of roughly the same number of conditions. However, there were changes in the After-minus-Before model and the post-intervention trend model. The After-minus-Before model predicted fewer conditions in Experiment 3; it no longer predicted Condition E and H. In Condition E, although the average causal strength was predicted by the After-minus-Before model, it appears that there was bimodality (both positive and negative judgments), so likely the After-minus-Before may not be the best explanation.

At the same time, the Post Trend model predicted all conditions but Condition C and some judgments in Condition E. It is possible that when the trials were spread out over a long timeframe, participants tended to focus more on more recent post-trend days when judging causal judgments. However, the reason that they focused on the recent days does not seem to be because they had better memory for those days; their memories appear to be fairly good for all 14 days. We will discuss this point more in the general discussion.

## 5. General discussion

The current research investigated how people assess cause-effect relationships in interrupted time series settings. In Experiments 1 and 2, we systematically tested a whole variety of interrupted time series conditions in the short timeframe. In Experiment 3, we investigated how people learn causality from interrupted time series in a long timeframe that mimicked real-life experience by presenting one trial per day for 14 days.

In the general discussion, the first 4 sections cover the 4 key findings: deviations from ITSA, that none of the theories could account for all of the findings, the minimal influence of presentation formats, and the minimal influence of presenting the data over a long timeframe. We then discuss the role of memory of individual events in causal judgment. Finally, we discuss some relations to other previous research not already mentioned and potential improvements in the dependent measures for future research.

### 5.1. Deviations from ITSA

The first key finding was that across all 3 studies, and within each of the 3 studies, there were many judgments that did not align with ITSA (though also some that did align with ITSA).

#### 5.1.1. Condition B: Pre-intervention slope

In Condition B, there was a positive (or negative) pre-intervention slope but no level change or slope change following the intervention, so ITSA infers no causal relationship. In all 3 experiments, most judgments were above 0 even though they should have been 0. The only exception was that in



Experiment 2, the predictive strength judgments were close to 0 and in line with ITSA. These findings are mostly consistent with the previous (White, 2015; Figure 1C) and can be explained by both the After-minus-Before model and the Post Trend model.

It is possible that when answering the causal strength question, when choosing that the ‘intervention caused the outcome to get higher’, what they really meant was simply that the outcome got higher, not that the intervention *caused* it to get higher. However, the participants also usually gave positive causal assessments for the future use question, which is not subject to this concern.

More broadly, Condition B can be viewed as a type of ‘illusory causation’—a situation in which participants infer a causal relation despite there not really being one—which has traditionally been viewed as an important error in causal judgment (e.g., Matute et al., 2015).

### 5.1.2. Condition E: Level change (incongruent)

In Condition E, the pre- and post-trends are negative but the level change is positive (or flipped), so ITSA infers a positive (or negative) causal relation. In all 3 experiments, most judgments were not in line with ITSA and the judgments were not consistent across experiments. More broadly, some of the histograms had high variability or bimodality (the histograms are available on OSF), suggesting that participants had multiple ways of thinking about the data in this condition. Additionally, participants’ memory of the trials around the level change (around Trial 8) was more deviated from the true values, compared to other conditions, which also indicates that participants were uncertain about the level change.

In Experiments 1 and 3 some judgments were negative, which is consistent with the post-intervention trend model. In Experiments 1 and 2 some judgments were close to 0, which is consistent with the After-minus-Before model. Another possibility is that participants were aware of both the positive level change (ITSA) and the negative post-intervention slope (or flipped), and gave judgments near 0 because of the opposing views. In Experiment 3, there was bimodality; some negative judgments and some positive ones.

One likely reason for the wide variety of responses in Condition E is the ‘incongruence’ between the pre and post-intervention slopes and the level change. If one does not account for the pre-intervention slope in incongruent situations (e.g., and instead just looks at After-minus-Before) they would make incorrect judgments. Or if one tends to judge causality based on local changes (change from one day to the next) instead of global changes, they would find that when they took the medicine, the symptom got worse from one day to the next, even though the symptom improved on the first day of taking the medicine compared to the day before. Disagreement with ITSA was also common in the other 2 incongruent conditions.

### 5.1.3. Condition H: Slope change (incongruent)

Condition H has a positive slope change and a negative pre-intervention slope (or flipped). In Condition H, participants tended to correctly infer a positive influence (or a negative influence for the flipped version); however, their inferences were quite weak compared to Condition F and G, even though all 3 conditions had exactly the same ITSA regression coefficient for the slope change (and all had very significant *p*-values for slope change). It is possible that some participants’ judgments were affected by similar mean levels in the pre- and post-intervention periods. Alternatively, it is possible that some participants were using the Post Trend heuristic; the post-intervention slope in Condition H was smaller than in the other 2 conditions.

### 5.1.4. Condition I: Slope change (maintain—also incongruent)

We called Condition I the ‘slope maintain’ condition because the post-intervention slope was 0 and the outcome stayed at the same level after the intervention was introduced. At the same time, the slope change according to ITSA is positive, but the pre-intervention slope is negative (or flipped). In Condition I, the judgments were close to 0 in all experiments. The most reasonable explanation is that participants were biased by the flat post-intervention trend and thought the intervention was ineffective,

which also fits with the Post Trend model. The judgments were also spread over the scale, which means there was considerable uncertainty when judging Condition I or potentially multiple types of responses. In [Section 5.6.1](#), we also discuss how alternative dependent measures may be more appropriate for Condition I.

### 5.1.5. Comparisons across conditions

According to ITSA, participants should make the same judgments in Conditions C and E, and also in Conditions F, H, and I. However, we found that their judgments were higher for C than E, and for F than H and I. Conditions E, H, and I, all involve cases when the pre-intervention slope was ‘incongruent’ with either the level change, or incongruent with the slope change, suggesting that incongruence impairs accurate judgment according to ITSA.

## 5.2. Comparing theories

The second key finding was that, though all of the models can account for a fair number of the results, none of the theories are *sufficient* to account for all of the results, and all of the theories are *necessary* to account for all of the results ([Figures 2 and 4](#)).

ITSA is necessary to explain the 14.6% of participants in Experiment 1, and the judgments in Condition B in Experiment 2. At the same time, there were also many deviations from ITSA, as described in the prior section.

After-minus-Before and Post-Trend are both necessary and insufficient to explain the judgments in Conditions E and I. For Condition E, After-minus-Before is necessary to explain the judgments near 0 in Experiment 2; however, Post-Trend is necessary to explain the negative judgments in Experiments 1 and 3. For Condition I, After-minus-Before is necessary to explain the negative judgments in Group 1 in Experiment 1; however, Post-Trend is necessary to explain the close-to-zero judgments in Group 3 in Experiment 1 and the close-to-zero judgments in Experiments 2 and 3.

It is possible that there are yet other strategies that would explain groups of participants better than these strategies. Alternatively, we think it is plausible that an individual person may primarily have one way to approach this problem, but when confronted with certain patterns may decide to answer in a different way. For example, in Group 3 in Experiment 1, participants for the most part answered in a way that is similar to the Post Trend model, but for Condition C in which there is just a level change but no pre-intervention slope nor a post-intervention slope, they still said that the intervention made a difference. We think that it is possible that these people were largely focusing on the Post Trend, but when confronted with the level change condition, realized that there are other ways to view the situation and that a level change could also be evidence of a causal influence even without any post trend. People may appeal to or notice different types of causal influence when observing different sorts of graphs or causal interactions, somewhat similar to the idea of causal pluralism (Wolff, 2014).

## 5.3. Presentation formats

The third key finding was that the presentation formats we tested had little influence on participants’ judgments. In particular, we tested 4 different presentation formats in Experiment 1; the only significant difference was in Condition E. In this condition participants in the trial-by-trial formats seemed to focus on the post-intervention trend; because the post-intervention trend was inconsistent with the level change, their inferences were in the wrong direction. In contrast, participants in the graph formats gave inferences on average near 0. This could be due to focusing on the mean levels of pre- and post-intervention periods, which was close to 0, or noticing the conflict between the level change and the post-intervention slope. The average inferences in both the trial-by-trial and graph conditions were wrong according to ITSA in that they failed to notice the level change, but the judgments in the trial-by-trial conditions were worse.

This finding, that the judgments in the trial-by-trial conditions were worse than in the graph conditions, at first glance seems opposite to what Soo and Rottman (2018, 2020) found, which is that the trial-by-trial conditions led to better judgments, and is the opposite of our hypothesis. However, in hindsight, we believe that there is a fairly straightforward explanation. Soo and Rottman argued that trial-by-trial presentations prompted participants to focus more on the local changes in both  $X$  and  $Y$  from one time to the next than the overall trends of  $X$  and  $Y$  (Figure 1D). In that study, the local changes in both  $X$  and  $Y$  helped uncover the true causal relation, whereas the overall trends were misleading. Furthermore, because  $X$  was continuous, there were many changes in  $X$ , which offered many opportunities to learn the local relation between  $X$  and  $Y$ .

However, in the current study, the cause is a binary variable and there was only one time that the cause changed (from Trial 7 to Trial 8). From the perspective of Soo and Rottman (2018), this provides only one opportunity to learn that in Condition E, an increase in  $X$  causes an increase in  $Y$ . In sum, Soo and Rottman proposed that trial-by-trial presentations make people notice local changes, and in Condition E, the local changes provide fairly little evidence about the level change. In contrast, when looking at the graphs, which promote noticing global patterns, it is probably more obvious that there are 2 things going on: the decreasing slope and the level change.

In sum, the critical difference between Condition E and Soo and Rottman's (2018, 2020) studies, is that in those studies the local changes provide the most helpful evidence for uncovering the causal relation, but in Condition E the global pattern provides the most helpful evidence for uncovering the causal relation. For all the other conditions in the current study, there is not such a clear distinction between local versus global as there is for Condition E, so it makes more sense why the format did not matter for those other conditions.<sup>11</sup> A natural future direction, particularly for the most problematic conditions like B, E, and I, is to test the influence of multiple interventions (e.g., starting, stopping, and starting the medicine, or increasing the dosage a few times) as opposed to just one (e.g., starting the medicine after 7 days). According to Soo and Rottman's account, having multiple interruptions in the time series would provide more local opportunities to learn the causal influence of the intervention, which could lead to improved judgments.

#### 5.4. Short versus long timeframes and theoretical models

The fourth key finding was that when participants experienced the data spread out over 2 weeks, their memories of the time series trends were fairly accurate, and their judgments were quite similar to the conditions in which they experienced the events presented rapidly or when shown graphs of the data.

In the introduction, and the introduction to Experiment 3, we raised multiple hypotheses about how learning and judgment might change for the long timeframe compared to the short, presuming that the long timeframe might have higher memory demands. Experiments 2 and 3 used identical designs and stimuli, except that one was conducted in the short timeframe (within a few minutes) but the other was in the long timeframe (over 2 weeks). It is not appropriate to formally compare these 2 studies; they had different samples of participants, and participants were not randomly assigned to the short versus long timeframes. Here we speculate about potential differences, but these should be interpreted with caution.

First, we hypothesized people may not be able to remember the earlier events in the long timeframe. In reality, Figure 7 shows that the averages of participants' memories were quite similar, in the short and long timeframes. The regression analyses in Table 2 that applied ITSA to participants' memories did reveal some inaccuracies in the memories in Experiment 3 that were not present in Experiment 2. However, most of these involved remembering a level change or slope change that did not exist. With regards to the pre-intervention slope, participants correctly remembered whether there was or was not a pre-intervention slope in 5 of the conditions. In Condition F, participants inferred a pre-intervention

<sup>11</sup>Perhaps for Condition I, it would also make sense for there to be an influence of format; according to local changes the intervention does not produce a change, but according to the global pattern the intervention influenced the slope of the outcome. Condition I will be discussed more below.

slope that did not exist, but the statistical result was quite weak ( $p = 0.018$ ). In sum, there is not much evidence that participants had an especially hard time remembering the pre-intervention events.

Second, we hypothesized that people in the long timeframe might resort to focusing on the post-intervention trend since it occurred most recently. There is some evidence of this. The Post-Intervention Trend Model explained more of the conditions for Experiment 3 (long timeframe) whereas the After-minus-Before explained more in Experiment 2 (short timeframe). The After-minus-Before model also was the best-fit model for most participants in Experiment 1, which was also a short timeframe study.

Third, the long timeframe could lead people toward using an associative-learning process such as RW, since they require very minimal memory resources. This would have predicted that participants would give negative judgments for Condition I. In contrast, similar to prior studies, participants continued to give judgments near 0 for Condition I, which is only consistent with the Post-Intervention Trend Model, not RW (nor After-minus-Before).

In conclusion, there were no radical differences between the judgments in the long versus short timeframe conditions, though there are some potential differences such as in Condition E. Furthermore, in both the long and the short timeframe, no single theory can account for all of the results, suggesting that people may use different reasoning or judgment processes in different situations, or that perhaps there is another explanation aside from the ones we put forth. Additionally, in Condition E in Experiment 3, there is some evidence of a bimodal distribution, suggesting that different participants may be thinking about this condition in different ways.

### ***5.5. Causal judgments, memory of the experiences, and potential interventions to improve causal judgments***

These studies provide multiple implications for the role of memory in causal judgments, and memory for the events more generally.

What did the participants remember about the events? Overall the memories of the outcomes of the 14 events were quite accurate, both in the short and long timeframe. The fairly high degree of accuracy may seem surprising, especially in the long timeframe. However, the relatively high accuracy does not mean that participants had accurate episodic memories for each of the 14 events. We believe that it is more likely that they developed a memory of the ‘gist’ or the global pattern (e.g., the outcome was flat prior to the intervention and then started increasing, or the outcome was decreasing before the intervention, but then started increasing).

At the same time, the fact that participants made fairly accurate predictions of the outcome for the upcoming trial suggests that during learning they had a relatively accurate memory of not just the overall pattern (e.g., flat and then increasing), but also the magnitude of the outcome. One plausible explanation is that they have a fairly accurate memory of the most recent outcome (e.g., 55, Bornstein et al., 2017), and the current trend (e.g., it has been increasing about 4 points per day after the intervention), and could use these two memories to predict the next outcome (e.g.,  $55 + 4 = 59$ ).

Overall, the ability to fairly accurately recall the patterns of events over many days could be an extremely useful foundation for causal inference. For example, people could use the memories of these patterns for making personal causal inferences, or for when making shared decisions (e.g., discussing whether a medication is working with a physician). And furthermore, the ability to fairly accurately predict upcoming daily events presumably helps people with planning and adapting if the events do not turn out as expected. If we had discovered that people’s memories of the patterns were very poor, this would have called for various kinds of recordings (e.g., daily diaries, passive monitoring systems) to aid causal inference, but since people’s memories are fairly accurate it seems that recordings are not necessary for the situations studied here (not necessarily all situations).

Yet despite the fairly accurate memories, some judgments deviated from ITSA. In the 2 short timeframe experiments, we found that participants made some inaccurate causal judgments, even though the memory demands were very low. In particular, the 2 graph formats in Experiment 1 did not require participants to use any memory, but participants still made some errors in causal judgments.

In Experiment 3, for which the trials were spaced out once per day, the memory demand should have been much greater, but we found that the judgments were similar in the short timeframe. In sum, the participants made systematic errors in causal judgments regardless of the memory demands of tasks, which suggests that the errors are due to faulty reasoning or judgment process rather than problems with learning or memory. This suggests that educational interventions tailored to critical reasoning about causality may be able to improve such judgments. Alternatively, in shared decision-making situations (e.g., discussions with a physician), it may make sense to first talk about the overall pattern of experiences and then to reason together about the inferences that can be drawn from the pattern, with guidance provided by the physician to overcome the reasoning errors.

## 5.6. Relations to other literature

### 5.6.1. Wolff's force dynamics model, and the phrasing of the dependent measures

Wolff's force dynamics model (2007) was not designed specifically for interrupted time series situations, but provides an interesting alternative way to think about the conditions studied here. The force dynamics model was designed to explain how people think about different notions of causality, and specifically to explain different concepts such as 'cause', 'enable'/'help', 'prevent', and 'despite'. The model proposes that people represent all sorts of causation, including 'nonphysical causation' (e.g., 'vitamin B enables the body to digest food', Wolff, 2007, p. 88) in a similar way to physical causation involving forces acting on objects (e.g., 'the wind enabled the boat to cross the Atlantic'). The model proposes that there are 2 entities, the 'affecter'/cause (e.g., wind), the patient (e.g., boat), and the interaction of the affecter on the patient leads to a particular end state or outcome of the patient (e.g., cross the Atlantic).

The model proposes that when understanding causal relations, 3 aspects need to be considered. The first is the tendency of the patient for the end state (e.g., whether the boat was already going to cross the Atlantic, the wind just helped it get there faster, or if it would not have succeeded without the wind). In the case of interrupted time series, we propose that the tendency of the patient for the end state can be understood as the pre-intervention trend of the outcome (e.g., whether the symptom was increasing before starting the medication). The second is whether the force of the affecter concurs with the force of the patient. For the boat example, in order to be an 'enable' situation, the wind must be pushing the boat in the same direction as the boat was moving on its own power. With regard to interrupted time series, we propose that what we have been calling congruency is similar to concordance. The third is the 'progress toward the end state', or really whether the end state was achieved (e.g., whether the boat eventually crossed the Atlantic). In the interrupted time series situations we studied, there was not a binary end state (e.g., the situation did not involve a symptom crossing a threshold of severity), so it is not exactly the same, but it could be viewed as similar to whether the outcome eventually got higher.

Using Wolff's categories, we propose that conditions C and F are cases of 'cause', I is a case of 'prevent', H could be either 'cause to go up' or 'prevent from going down', D and G are cases of 'enable/help', and E is a case of 'despite'. Conditions A and B, which are noncausal, do not have a corresponding word. Thus, Wolff's dynamics model distinguishes the 'no influence' conditions (A and B) from the conditions that do have a causal influence, similar to ITSA (conditions C–I), but also makes additional distinctions not made by ITSA.

This conceptualization raises the possibility of modifying the dependent measures for future studies. For example, for Condition E, participants tended to give neutral or negative judgments despite there being a positive level change. It is possible that they gave negative responses from a practical perspective that even if the medicine increases the outcome, it only does so temporarily, or equally, only delays how long it will take for the outcome to reach a certain low level. This may not be viewed as sufficient to use a medicine. However, it is possible that participants would have agreed with a statement like 'Despite the positive influence of the medicine, the outcome continued to decrease'. Likewise, for Condition I in which the initial slope becomes flat after the intervention, they may have agreed with a



statement like ‘The medicine prevented the outcome from decreasing’.<sup>12</sup> In sum, using questions that are arguably more linguistically appropriate may have revealed judgments more in line with ITSA than what we found in the causal strength judgments that only used the verb ‘cause’.

At the same time, the predictive strength question is not subject to this critique, and these judgments also deviated considerably from ITSA. The judgments for Condition B also deviated from ITSA despite the question being linguistically appropriate. In sum, asking alternative questions would likely reveal additional nuances in the conceptualization of causality, though it is not clear that participants’ judgments would agree with ITSA.

### 5.6.2. The counterfactual simulation model

Another relevant model of causal reasoning, the counterfactual simulation model (CSM), proposes that people make causal judgments by comparing what actually happened with what would have happened in a counterfactual scenario (Gerstenberg et al., 2021; Wu et al., 2022). According to CSM, the extent to which a candidate actually caused an outcome is based on how certain one feels that the outcome would not have happened, or would have happened in a different way, if the candidate had been absent. Because a counterfactual simulation requires people to compare the current level of the outcome with what would have happened according to the existing pre-trend and no intervention, the CSM model makes the same judgments as ITSA. However, in several conditions, participants failed to make judgments in line with the CSM / ITSA.

### 5.6.3. Controlling for variables

There is a long history of research on whether and when people ‘control for variables’. Some of this research has examined whether people control for third variables when assessing the relation between 2 variables (e.g., Spellman, 1996; Vanderpe and Houwer, 2006; Waldmann, 2000; Waldmann and Hagmayer, 2001). Another set of research has examined how people learn to control for third variables when designing a study to test the relation between 2 variables—specifically designing unconfounded studies (e.g., Chen and Klahr, 1999; Kuhn and Dean, 2005; Schwichow et al., 2016; van der Graaf et al., 2015).

All of this research is related to the current study in that in the current study it is important to control for temporal trends that occur prior to the intervention. At the same time, a fundamental difference between those studies and the current one is that in those studies the alternative cause was directly observed, not a latent factor like time, and in those studies, the alternative cause was binary making it easier to control for. In sum, the current work can be thought of as an extension of this long line of research on controlling for variables but instead of controlling for variables that are present versus absent in individual instances, it involves understanding processes unfolding over time (Grotzer et al., 2013) which is all the more relevant when the patterns of events unfold over weeks.

### 5.6.4. Nonlinear systems

The current study investigated ITSA with a linear system; at the same time, there are many systems that are nonlinear and participants might have believed that our cover story about a biological system might be nonlinear. In particular, participants might have believed that there could be asymptotes that prevent the outcomes from getting higher or lower. Gong and Bramley (2022) studied how people reason about a system with a continuous cause that undergoes one big change (similar to our ‘interventions’) and a continuous effect that has upper and lower asymptotes. Their participants gave stronger judgments when the effect reached an asymptote faster (somewhat similar to a strong slope change). Their participants were also more sensitive to changes that occurred close to asymptotes. For example, they gave stronger causal judgments when the effect dropped from 24 to 8, close to the boundary of 0,

<sup>12</sup>We considered using more nuanced dependent measures for these studies, but decided not to do so because it would have been a very complicated measure; it would have involved asking four questions (cause, prevent, enable, despite), each for two outcomes (got higher, got lower), as well as a no influence option. But we believe that this approach would be fruitful for future research.



then when it dropped from 48 to 24. Because we were concerned that participants might believe the stimuli had asymptotes, in the cover stories for Experiments 2 and 3 we conveyed to participants that the outcome could go higher or lower than the range of the graph, and the graph only showed the range of 30 to 90.

Additionally, participants might believe that the system is homeostatic in which the system tends to come back toward some normal value (Rehder et al., 2022). Because the data we presented to participants were so strongly linear, it seems unlikely to us that participants believed them to be homeostatic or asymptotic. However, if people have very different assumptions about how the processes work than the ITSA process that we assumed, then ITSA would not be the ‘normative’ model that people should be compared against. These other types of systems should be investigated more in future research.

## 6. Conclusions

The current study found that though people’s causal judgments about interrupted time series scenarios sometimes align with formal ITSA, in many conditions they do not. The reason for the judgments that do not align with ITSA does not seem to do with the format—the format has fairly little influence except in one condition. These issues persist, but are not exaggerated when the events are experienced in a more realistic timeframe, once a day over 2 weeks. Furthermore, participants’ memories were fairly accurate in both short and long timeframes, suggesting that these errors are likely due to reasoning and judgment, not learning and memory. Still, none of the existing theories could explain all of the findings, suggesting that people use multiple different ways to think about ITSA situations. Future research could focus on creating interventions for getting people to more accurately interpret interrupted time series data. Another project for future research is to examine the accuracy of judgments with 2 or more potential causes or effects, in which case memory may become more of a limiting factor.

**Data availability statement.** The registration plans and data are available here: <https://osf.io/nqb8c/>.

**Funding statement.** This research was funded by the National Science Foundation (grant 1651330).

**Competing interest.** The authors have no known conflicts of interest to disclose.

## References

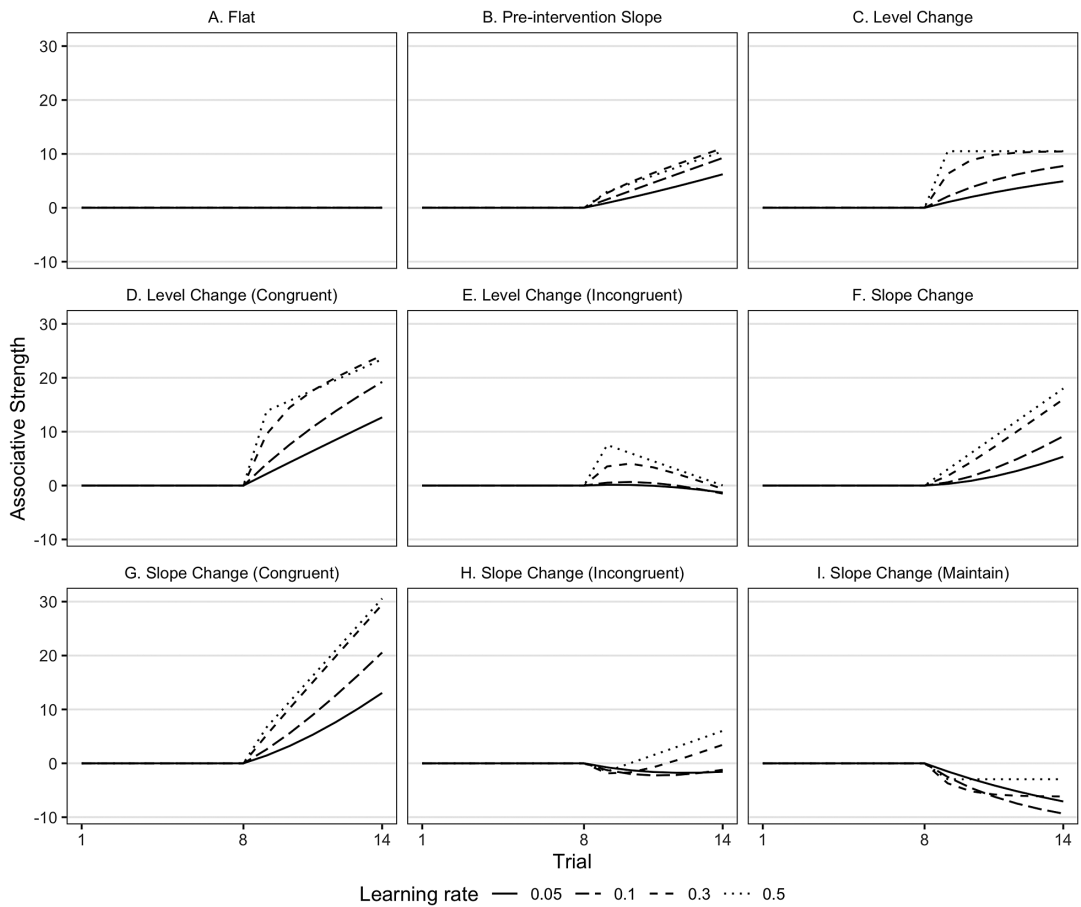
- Barlow, D. H. & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, 12(2), 199–210. <https://doi.org/10.1901/jaba.1979.12-199>
- Bernal, J. L., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology*, 46(1), 348–355. <https://doi.org/10.1093/ije/dyw098>
- Bishara, A. J., Peller, J., & Galuska, C. M. (2021). Misjudgment of interrupted time-series graphs due to serial dependence: Replication of Matyas and Greenwood (1990). *Judgment and Decision Making*, 16(3), 22.
- 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(1), 1. <https://doi.org/10.1038/ncomms15958>
- 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), 1880–1910. <https://doi.org/10.1037/xlm0000548>
- Bramley, N. R., Mayrhofer, R., Gerstenberg, T., & Lagnado, D. A. (2017). Causal learning from interventions and dynamics in continuous time. In *Proceedings of the 39th annual meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Caddick, Z. A., & Rottman, B. M. (2019). Politically motivated causal evaluations of economic performance. In *Proceedings of the 41st annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70(5), 1098–1120. <https://doi.org/10.1111/1467-8624.00081>
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405. <https://doi.org/10.1037/0033-295X.104.2.367>
- Collins, A. G. E. (2018). The tortoise and the hare: Interactions between reinforcement learning and working memory. *Journal of Cognitive Neuroscience*, 30(10), 1422–1432. [https://doi.org/10.1162/jocn\\_a\\_01238](https://doi.org/10.1162/jocn_a_01238)

- Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, 47(2), 109–121. [https://doi.org/10.1016/S0022-2496\(02\)00016-0](https://doi.org/10.1016/S0022-2496(02)00016-0)
- Davis, Z. J., Bramley, N. R., & Rehder, B. (2020). Causal structure learning in continuous systems. *Frontiers in Psychology*, 11, 244. <https://www.frontiersin.org/article/10.3389/fpsyg.2020.00244>
- Deprospero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, 12(4), 573–579. <https://doi.org/10.1901/jaba.1979.12-573>
- Derringer, C., & Rottman, B. M. (2018). How people learn about causal influence when there are many possible causes: A model based on informative transitions. *Cognitive Psychology*, 102, 41–71. <https://doi.org/10.1016/j.cogpsych.2018.01.002>
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 128(5), 936–975. <https://doi.org/10.1037/rev0000281>
- Gong, T., & Bramley, N. R. (2022). Intuitions and perceptual constraints on causal learning from dynamics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44, 1455–1461. <https://escholarship.org/uc/item/1d63r60d>
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51(4), 334–384. <https://doi.org/10.1016/j.cogpsych.2005.05.004>
- Grotzer, T. A., Kamarainen, A. M., Tutwiler, M. S., Metcalf, S., & Dede, C. (2013). Learning to reason about ecosystems dynamics over time: The challenges of an event-based causal focus. *BioScience*, 63(4), 288–296. <https://doi.org/10.1525/bio.2013.63.4.9>
- Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis*, 13(4), 543–559. <https://doi.org/10.1901/jaba.1980.13-543>
- 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. <https://doi.org/10.1080/03640210701530755>
- Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, 11(2), 277–283. <https://doi.org/10.1901/jaba.1978.11-277>
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, 16(11), 866–870. <https://doi.org/10.1111/j.1467-9280.2005.01628.x>
- Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, 3(1), 184–195. <https://doi.org/10.2174/1874350101003010184>
- Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. A. (2012). A “SMART” design for building individualized treatment sequences. *Annual Review of Clinical Psychology*, 8(1), 21–48. <https://doi.org/10.1146/annurev-clinpsy-032511-143152>
- Matute, H., Blanco, F., & Diaz-Lago, M. (2019). Learning mechanisms underlying accurate and biased contingency judgments. *Journal of Experimental Psychology: Animal Learning and Cognition*, 45(4), 373–389. <https://doi.org/10.1037/xan0000222>
- Matute, H., Blanco, F., Yarritu, I., Diaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, 6, 888. <https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00888>
- Memory of Alfred Nobel . (2021). Answering causal questions using observational data (pp. 1–48). Stockholm: The Royal Swedish Academy of Sciences.
- Minford, A. M., MacDonald, A., & Littlewood, J. M. (1982). Food intolerance and food allergy in children: A review of 68 cases. *Archives of Disease in Childhood*, 57(10), 742–747. <https://doi.org/10.1136/adc.57.10.742>
- Rehder, B., Davis, Z. J., & Bramley, N. (2022). The paradox of time in dynamic causal systems. *Entropy*, 24(7), 7. <https://doi.org/10.3390/e24070863>
- Rescorla, R. A., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current research and theory* (Vol. 2). New York: Appleton-Century-Crofts.
- Rottman, B. M. (2016). Searching for the best cause: Roles of mechanism beliefs, autocorrelation, and exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(8), 1233–1256. <https://doi.org/10.1037/xlm0000244>
- Rottman, B. M., & Ahn, W. (2009). Causal learning about tolerance and sensitization. *Psychonomic Bulletin & Review*, 16(6), 1043–1049. <https://doi.org/10.3758/PBR.16.6.1043>
- Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, 64(1), 93–125. <https://doi.org/10.1016/j.cogpsych.2011.10.003>
- Rottman, B. M., Kominsky, J. F., & Keil, F. C. (2014). Children use temporal cues to learn causal directionality. *Cognitive Science*, 38(3), 489–513. <https://doi.org/10.1111/cogs.12070>
- Schwichow, M., Coker, S., Zimmerman, C., Höfler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, 39, 37–63. <https://doi.org/10.1016/j.dr.2015.12.001>
- Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology*, 60(3), 291–309. <https://doi.org/10.1080/17470210601000581>
- Soo, K. W., & Rottman, B. M. (2018). Causal strength induction from time series data. *Journal of Experimental Psychology: General*, 147(4), 485–513. <https://doi.org/10.1037/xge0000423>
- Soo, K. W., & Rottman, B. M. (2020). Distinguishing causation and correlation: Causal learning from time-series graphs with trends. *Cognition*, 195, 104079. <https://doi.org/10.1016/j.cognition.2019.104079>
- Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science*, 7(6), 337–342. <https://doi.org/10.1111/j.1467-9280.1996.tb00385.x>

- van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science*, 43(3), 381–400. <https://doi.org/10.1007/s11251-015-9344-y>
- Vandorpe, S., & Houwer, J. D. (2006). A comparison of cue competition in a simple and a complex design. *Acta Psychologica*, 122(3), 234–246. <https://doi.org/10.1016/j.actpsy.2005.11.003>
- Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1), 53–76. <https://doi.org/10.1037/0278-7393.26.1.53>
- Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, 82(1), 27–58. [https://doi.org/10.1016/S0010-0277\(01\)00141-X](https://doi.org/10.1016/S0010-0277(01)00141-X)
- Well, A. D., Boyce, S. J., Morris, R. K., Shinjo, M., & Chumbley, J. I. (1988). Prediction and judgment as indicators of sensitivity to covariation of continuous variables. *Memory & Cognition*, 16(3), 271–280. <https://doi.org/10.3758/BF03197760>
- White, P. A. (2015). Causal judgements about temporal sequences of events in single individuals. *Quarterly Journal of Experimental Psychology*, 68(11), 2149–2174. <https://doi.org/10.1080/17470218.2015.1009475>
- White, P. A. (2017). Causal judgments about empirical information in an interrupted time series design. *Quarterly Journal of Experimental Psychology*, 70(1), 18–35. <https://doi.org/10.1080/17470218.2015.1115886>
- Willett, C. L., & Rottman, B. M. (2020). Causal learning with two causes over weeks. In *Proceedings of the annual conference of the Cognitive Science Society*. Alexandria, VA: The National Science Foundation. <https://par.nsf.gov/biblio/10237620-causal-learning-two-causes-over-weeks>
- Willett, C. L., & Rottman, B. M. (2021). The accuracy of causal learning over long timeframes: An ecological momentary experiment approach. *Cognitive Science*, 45(7), e12985. <https://doi.org/10.1111/cogs.12985>
- Wilson, I. B., Schoen, C., Neuman, P., Strollo, M. K., Rogers, W. H., Chang, H., & Safran, D. G. (2007). Physician–patient communication about prescription medication nonadherence: A 50-state study of America’s seniors. *Journal of General Internal Medicine*, 22(1), 6–12. <https://doi.org/10.1007/s11606-006-0093-0>
- Wimmer, G. E., Li, J. K., Gorgolewski, K. J., & Poldrack, R. A., (2018). Reward learning over weeks versus minutes increases the neural representation of value in the human brain. *The Journal of Neuroscience*, 38(35), 7649–7666. <https://doi.org/10.1523/JNEUROSCI.0075-18.2018>
- Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, 136(1), 82–111. <https://doi.org/10.1037/0096-3445.136.1.82>
- Wolff, P. (2014). Causal pluralism and force dynamics. In B. Copley & F. Martin (Eds.), *Causation in grammatical structures* (pp. 100–119). Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199672073.003.0005>
- Wu, S. A., Sridhar, S., & Gerstenberg, T. (2022). That was close! A counterfactual simulation model of causal judgments about decisions [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/7uwbf>
- Zhang, Y., & Rottman, B. (2021). Causal learning with delays up to 21 hours. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43), 2766–2772. <https://escholarship.org/uc/item/8w78273f>

## A. Appendix A

We simulated the Rescorla–Wagner model (RW) for each interrupted time series condition. In our simulations, we moved the outcomes down by a constant so that the level of outcome was always 0 on Day 1, which speeds up learning of the background cause. We also tested high and low learning rates. [Figure A1](#) shows the associative strength of the intervention for 14 trials with 4 different learning rates. No learning occurs for Days 1–7 since the cause is absent on those days.



**Figure A1.** Simulations of Rescorla–Wagner.

## B. Appendix B

**Table B1.** Inferential statistics for Experiment 1.

| Statistics                    | Causal strength          |         |       |        |       | Future use               |         |       |        |       |
|-------------------------------|--------------------------|---------|-------|--------|-------|--------------------------|---------|-------|--------|-------|
|                               | <i>t</i> -test against 0 |         |       |        | ANOVA | <i>t</i> -test against 0 |         |       |        | ANOVA |
|                               | Static                   | Dynamic | Bar   | Number |       | Static                   | Dynamic | Bar   | Number |       |
| A. Flat                       |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | .62                      | .476    | .475  | .629   | .805  | <.001                    | <.001   | <.001 | <.001  | .316  |
| BF                            | .13                      | .14     | .14   | .13    | .02   | >100                     | >100    | >100  | >100   | .05   |
| Effect size                   | .05                      | .07     | .07   | .05    | <.001 | .54                      | .69     | .75   | .55    | .01   |
| B. Pre-intervention slope     |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | <.001                    | <.001   | <.001 | <.001  | .088  | <.001                    | <.001   | <.001 | <.001  | .018  |
| BF                            | >100                     | >100    | >100  | >100   | .19   | >100                     | >100    | >100  | >100   | .97   |
| Effect size                   | .65                      | .91     | .89   | 1.09   | .02   | .49                      | .99     | .89   | 1.04   | .03   |
| C. Level change               |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | <.001                    | <.001   | <.001 | <.001  | .832  | <.001                    | <.001   | <.001 | <.001  | .48   |
| BF                            | >100                     | >100    | >100  | >100   | .02   | >100                     | >100    | >100  | >100   | .03   |
| Effect size                   | .83                      | 1.03    | 1.04  | .94    | <.001 | .78                      | 1.09    | .76   | .72    | .01   |
| D. Level change (congruent)   |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | <.001                    | <.001   | <.001 | <.001  | .033  | <.001                    | <.001   | <.001 | <.001  | .244  |
| BF                            | >100                     | >100    | >100  | >100   | .53   | >100                     | >100    | >100  | >100   | .07   |
| Effect size                   | 1.01                     | 2.41    | 1.61  | 1.05   | .02   | .93                      | 1.58    | 1.48  | 1.44   | .01   |
| E. Level change (incongruent) |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | .119                     | .077    | <.001 | <.001  | <.001 | .195                     | .035    | <.001 | <.001  | <.001 |
| BF                            | .37                      | .52     | >100  | >100   | 88.63 | .26                      | .99     | >100  | >100   | 38.67 |
| Effect size                   | .16                      | .18     | .67   | .51    | .05   | .13                      | .22     | .77   | .53    | .05   |
| F. Slope change               |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | <.001                    | <.001   | <.001 | <.001  | .078  | <.001                    | <.001   | <.001 | <.001  | .103  |
| BF                            | >100                     | >100    | >100  | >100   | .22   | >100                     | >100    | >100  | >100   | .17   |
| Effect size                   | 1.34                     | 1.93    | 1.19  | 1.24   | .02   | 1.37                     | 1.68    | 1.24  | 1.02   | .02   |
| G. Slope change (congruent)   |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | <.001                    | <.001   | <.001 | <.001  | .032  | <.001                    | <.001   | <.001 | <.001  | .317  |
| BF                            | >100                     | >100    | >100  | >100   | .55   | >100                     | >100    | >100  | >100   | .05   |
| Effect size                   | .96                      | 1.98    | 1.67  | 1.43   | .02   | .82                      | 1.31    | 1.05  | 1.15   | .01   |
| H. Slope change (incongruent) |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | <.001                    | <.001   | <.001 | <.001  | .995  | <.001                    | <.001   | <.001 | <.001  | .895  |
| BF                            | >100                     | >100    | >100  | >100   | .01   | >100                     | >100    | >100  | >100   | .01   |
| Effect size                   | .48                      | .52     | .58   | .46    | <.001 | .48                      | .58     | .57   | .56    | <.001 |
| I. Slope change (maintain)    |                          |         |       |        |       |                          |         |       |        |       |
| <i>p</i>                      | .265                     | .576    | .797  | .854   | .579  | .265                     | .008    | .012  | .619   | .017  |
| BF                            | .21                      | .13     | .11   | .11    | .03   | .21                      | 3.5     | 2.45  | .13    | 1     |
| Effect size                   | .11                      | .06     | .03   | .02    | .01   | .11                      | .27     | .26   | .05    | .03   |

Note: For effect size, we calculated Cohen's *d* for all *t*-tests and  $R^2$  for ANOVAs.