Causal Learning With Interrupted Time Series

Yiwen Zhang (yiwenzhang@pitt.edu)

Benjamin M. Rottman (rottman@pitt.edu)
Department of Psychology, University of Pittsburgh
Pittsburgh, PA 15260 USA

Abstract

Interrupted time series analysis (ITSA) is a statistical procedure that evaluates whether an intervention causes a change in the intercept and/or slope of the time series. However, very little research has accessed causal learning in interrupted time series situations. We systematically investigated whether people are able to learn causal influences from a process akin to ITSA, and compared four different presentation formats of stimuli. We found that participants' judgments agreed with ITSA in cases in which the pre-intervention slope is zero or in the same direction as the changes in intercept or slope. However, participants had considerable difficulty controlling for pre-intervention slope when it is in the opposite direction of the changes in intercept or slope. The presentation formats didn't affect judgments in most cases, but did in one. We discuss these results in terms of two potential heuristics that people might use aside from a process akin to ITSA.

Keywords: causal learning; interrupted time series analysis; presentation formats

Introduction

For assessing cause-effect relationships in time series data, randomised controlled experiments are usually not available. For example, when a patient wants to test the efficacy of a new medicine for treating their depression, they can only track their depression from before to after taking the medicine to tell whether the depression has been improved by the medicine. Or similarly, when a country makes a change in the economy (e.g., lowers the interest rate) and wants to look for changes in economic variables, it makes sense to track the trend before and after the change to see if there has been a change in the time series. Importantly, in such situations, reasoners need to control for the temporal trend of the outcome.

Interrupted Time Series Analysis

Interrupted time series analysis (hereafter ITSA) is a useful method to assess the influence of an intervention within time series data (Hartmann et al., 1980). A simple ITSA is modeled with three components using a regression model:

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

where T is the time since the start of first observation; X is a dummy variable indicating whether or not the intervention has been conducted; β_0 indicates the intercept of the outcome when $T = 0$, β_1 indicates the pre-intervention slope of the time series, β_2 indicates the changes in the intercept from pre-intervention to post-intervention period, and β_3 indicates the changes in the slope from the pre-intervention to post-intervention period (Bernal, Cummins, & Gasparrini, 2016).

Causal Learning from Interrupted Time Series

Despite time series being an important sort of data for people to make causal inferences from, only a few studies have examined how people make causal inferences in interrupted time series situations. One study by White (2015) has found that people are not sensitive to the pre-intervention trends in the time series. Participants were shown time series data which contained a period in which the data increased and were asked to judge the causal strength of an intervention. The intervention occurred either before, in the middle of, or after the increasing period. Surprisingly, in most cases, participants judged that the midway intervention was as effective as the intervention which happened prior to the increasing period. That is, they failed to understand that the trend was already increasing before the intervention, so any additional increase along the same trajectory could simply be a continuation of the prior trend. White proposed the 'after-minus-before' model to explain participants' logic, which involves simply comparing the mean level of the pre-intervention and post-intervention time periods, and ignoring the possibility of a trend that started before the intervention. However, White (2015, 2017) did find exceptions that were not captured by the after-minus-before model: (1) the intervention with immediate effects received higher causal strength judgments than the intervention with delayed effects; (2) with graphical presentation of stimuli, participants were somewhat able to control for the pre-intervention slope and gave lower causal strength judgements for interventions that occurred in the middle of the slope.

Though there has been little research specifically on interrupted time series situations, there have been some other studies on how people make causal inferences from time-series data (e.g., Rottman, 2016; Derringer & Rottman, 2016). The important distinction between these studies and the prior ones is that these involved multiple changes to the potential cause (among other differences). However, these studies have found two important things. First, people tend to focus on how the effect changes when the cause changes (i.e., changes in trends), which often allows people to control for temporal trends in the data, not merely what the correlation is between the cause and the effect (similar to the after-minus-before).

Furthermore, there is some evidence that, similar to White (2017), the presentation of the data, be it in a graph vs. numbers presented all at once vs. stimuli presented trial by trial can make different patterns more or less salient and affect the accuracy of participants' judgments. For example, Soo and Rottman (2020) found that dynamic presentations helped people accurately learn causal relationships by focusing on changes in the cause and effect, whereas static and numerical presentations led them to focus on the simple correlation and not account for trends.

Potential Theories

In the current research we investigated three potential theories for how people might make inferences from interrupted time series data. First, they might implement a process similar to formal ITSA, and look for changes in the intercept or slope after the intervention compared to before. Second, they might use the after-minus-before heuristic proposed by White (2015), which involves compare the mean of the data after the intervention vs. before.

We also propose a third theory which we call 'post-intervention trend' (or 'post. trend' for short) that involves simply focusing on the slope of the post-intervention trend. We initially came up with this theory from some participants' explanations for judgments in a pretest. The idea is that a positive (negative) post-intervention trend is interpreted as evidence that the intervention increases (decreases) the outcome. This sort of inference is clearly non-rational (e.g., a trend could have increased even more if the intervention did not occur). However, it could be understood from a perspective of feeling that something must be responsible for changes in the outcome, and repeated experiences of the cause could be responsible for repeated changes in the effect.

Current study

The main goal of the current research is to extend the research conducted by White (2015, 2017) by investigating interrupted time series situations, but in a more systematic way. White only investigated a limited set of cases, and did not actually use interrupted time series data analysis to generate the stimuli. In the current study we generated a spectrum of situations that either do or do not have pre-intervention slopes, and have various post-intervention changes to the intercept or to the slope. We also examined the effect of four presentation formats of the interrupted time series in order so that the findings are relevant to data presented in a summarized graphical forms and also data experienced sequentially over time.

Methods

Participants

402 participants were recruited on Mechanical Turk; the preregistraction said we would recruit 400 and 2 additional participants did the study without submitting the HIT. All participants had an overall HIT Approval Rate that is greater than or equal to 95%. The experiment lasted 10-15 minutes and participants were paid $2. This is our first study on this topic so we do not have estimates of effect size of potential effects. The study is within-subjects providing a fairly high degree of power.

Cover story

Participants were told to imagine that they work for a medicine company that is testing the efficacy of new medications. They reviewed 9 datasets, each with a different medicine (e.g., SNP27), and a different symptom (e.g. headache, back pain). Each depicted the data for a single patient over an initial week without and one week with the medicine.

Stimuli and Design

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

Data for Nine Time Series Conditions For the 9 interrupted time series conditions, see the first column in Figure 1. These 9 include situations with and without pre-intervention slopes, and with and without changes in the intercept or slope to systematically examine a broad variety of interrupted time series situations. We did not investigate situations that have changes to both the intercept and the slope. Table 1 shows the predictions made by the various theories.

Though some of the conditions are self-explanatory, a couple need to be explained. The B. Pre-Intervention Slope condition is similar to White's (2015) condition in which the intervention occurred in the middle of a slope.

The E. Slope Change (maintain) condition is similar to the D. Slope Change condition, in that they both involve a slope change, however for Condition E, the post-intervention slope is zero. According to ITSA participants should infer positive causal efficacy. According to after-minus-before they should infer negative, and according to post-intervention-trend they should not infer any influence of the cause (Table 1).

Conditions F-I involve either changes to the intercept or slope in addition to having a pre-intervention trend. We call F and G 'congruent' in that the change in the intercept or change in the slope is in the same direction as the pre-intervention slope. This means that all three theories make the same predictions for the congruent conditions. However, in Conditions H and I, the changes in intercept and slope are 'incongruent' with the initial slope (e.g., in H the intercept change is positive whereas the initial slope is negative). For Conditions H and I, the theories do not all make the same predictions; see Table 1.[1]

[1]The post-intervention mean in Conditions H and I were slightly lower than the pre-intervention mean but the difference is so small it might be hard for participants to notice even if trying. The reason that it was not exactly 0 was due to way we set up the same value of coefficients across conditions, though in future research it would also be valuable to make these exactly 0. This is why in Talbe 1 they are listed as slightly negative.

Table 1: ITSA coefficients, model predictions, and simplified empirical data results for the 9 time series condition

	ITSA Coefficients			Model Predictions			Empirical
Condition	Pre. Slope	Intercept Change	Slope Change	ITSA	After-before	Post. Trend	Data
A. Flat	0	0	0	0	0	0	0
B. Pre-intervention Slope	3.85	0	0	0	+	+	+
C. Intercept Change	0	+20	0	+	+	0	+
D. Slope Change	0	0	+3.08	+	+	+	+
E. Slope Change (Maintain)	-3.08	0	+3.08	+	-	0	0
F. Intercept Change (C. - PS)	3.85	+20	0	+	+	+	+
G. Slope Change (C. - PS)	3.85	0	+3.08	+	+	+	+
H. Intercept Change (I. - PS)	-3.85	+20	0	+	slightly -	-	-/0
I. Slope Change (I. - PS)	-3.85	0	+3.08	+	slightly -	+	+

Note: This table only presents coefficients for positive datasets.
C.-PS: congruent pre-intervention slope; I.-PS: incongruent pre-intervention slope;
Pre. Slope is short for Pre-intervention Slope; Post. Trend is short for Post-intervention Trend.

Except for the A. Flat condition, we included two parallel datasets for each time series condition. These parallel datasets simply involved flipping the Y axis, for generality. For all these other conditions (B-I), participants saw one version or the other (4 from the versions depicted in first column of Figure 1, and 4 from the flipped version) randomly selected. For data analysis, all judgments were reverse coded for the flipped versions and analyzed together.

The numbers for the datasets were generated in the following way, and always produced numbers between 0 and 100. First we created the baselines functions for the 9 time series conditions using the coefficients in Table) and Equation 1. We then then added pseudo-Gaussian noise to the baseline functions and rounded the datasets to be 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 stimuli, ITSA uncovered coefficients similar to those in Table 1, and all p-values for the non-zero coefficients were in the range of [10^{-9}, 10^{-7}]. Table 1 shows both the ITSA coefficients, and also the simplified model predictions; the ITSA model predicts a positive outcome whenever the intercept change or slope change are positive.

Four Presentation formats Figure 2 shows the four 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 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 which indicated the status of the intervention (medicine) on the left, and a bar (a narrow dot chart) with a dot which 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 participant from going through the observations too fast, we set a 1-second waiting time between each click in the dynamic graph and two trial-by-trial conditions.

In static graph format, the questions were shown at the same time with the graph. In the dynamic graph format, the questions were shown below the graph once all the 14 data points were revealed. In the two trial-by-trial formats, the questions were shown on a new page after the 14 trials.

Measures

After participants reviewed the observations in a dataset, they answered three questions about the influence of the medicine.

We measured the causal strength by asking participants "Did taking [medicine] cause the [symptom] to get better or worse?" on a scale from 1-9 scale: 1 (the medicine caused the symptom get much worse - higher), 5 (the medicine had no influence on the symptom), to 9 (the medicine caused the symptom get much better - lower).

We measured the 'future use' by asking participants "Do you think this patient should continue to take the medicine to treat the symptom?" on a 1-9 scale: 1 (should definitely stop taking the medicine), 5 (unsure whether to continue or to stop), to 9 (should definitely continue to take the medicine).

They also answered a free-response question: "Please explain how you answered the questions above." We do not analyze this question in this article.

Results

The analysis follows our pre-registered plan (https://osf.io/uzt37). Since we included two parallel datasets for each time series condition, we inversely coded the responses from those datasets with a positive pre-intervention slope in Condition B and positive changes in slope or intercept in Condition C to I, because when a medicine caused the level of a symptom to get worse (higher), that would correspond to a causal judgment below the mid point on the scale. We also inversely coded Condition A to make the judgements have the same meanings as other conditions. We then collapsed two parallel datasets

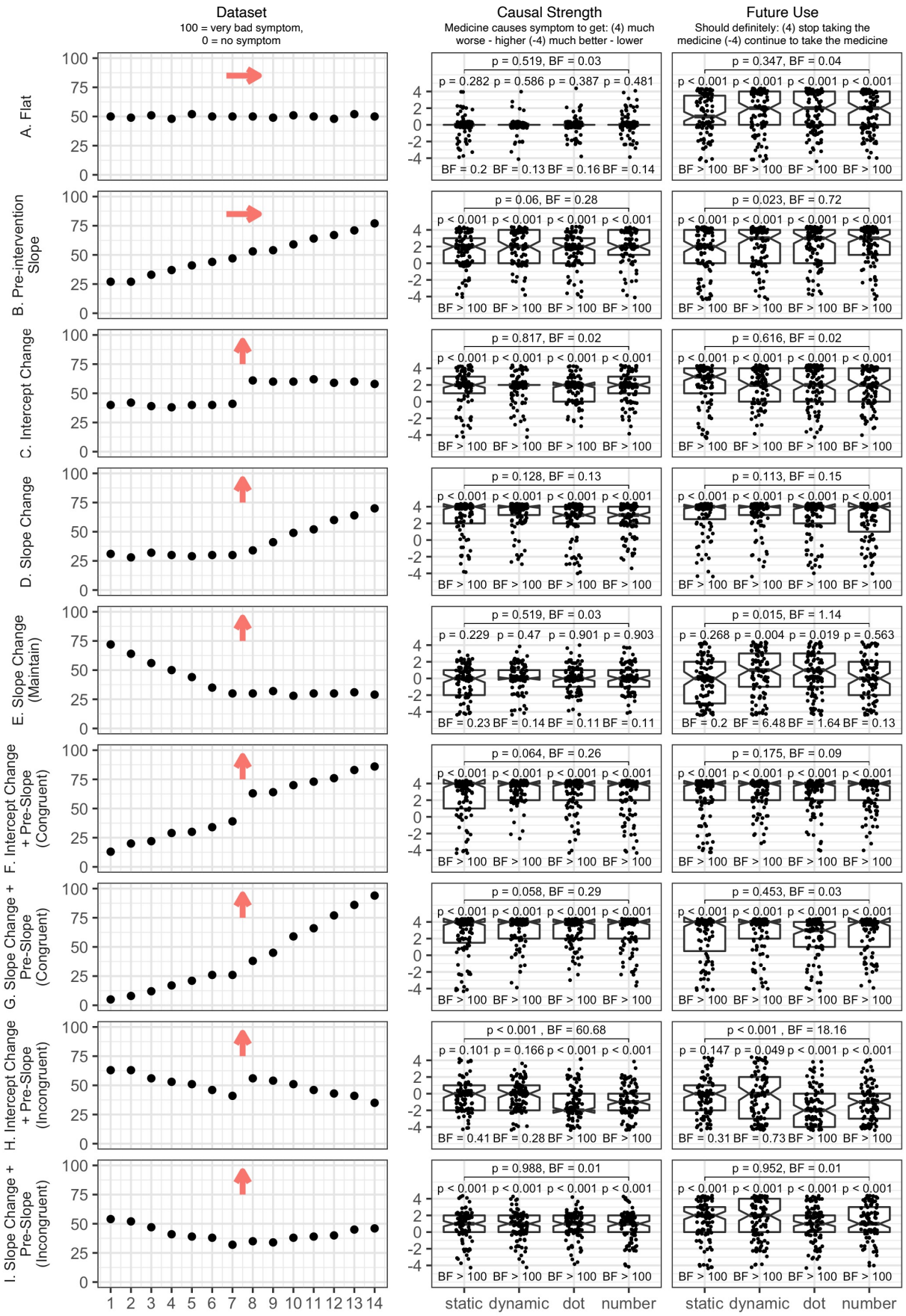

Figure 1: Stimuli and results. Column 1 shows example stimuli. The red arrow indicates the influence of intervention according to ITSA (→ = no influence; ↑ = positive influence). Columns 2 and 3 show a summary of the results in the four formats.

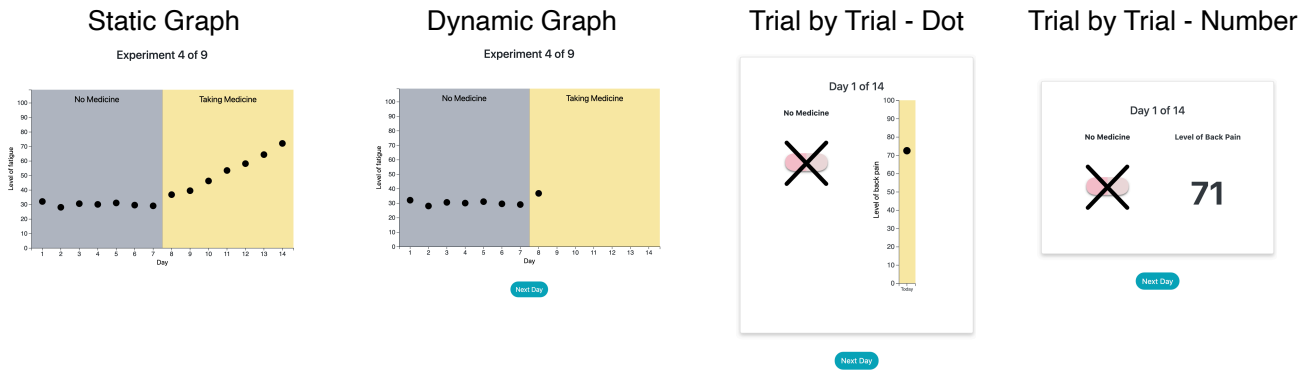

Figure 2: Screenshots of four presentation formats.

of each time series condition for data analysis. For ease of interpretation, we centered the measures around zero so that the ranges were [-4,4]. In Figure 1, we plotted the datasets in which changes in slope or intercept that make symptoms get worse/higher (positive datasets), and for the dependent measures, higher means the medicine caused the symptom to get worse.

For both two measures, we conducted two analysis. First, we conducted one sample t-test against zero for each time series condition and format to see if the judgments fit the ITSA predictions. Second, we conducted ANOVA to compare four presentation formats within a time series condition.

Figure 1 depicts all the results as well as inferential statistics. We provide p values and Bayes Factors (BFs, Kruschke, 2014). BFs less than 1 favor the null, and greater than 1 favor the alternative hypothesis. The red arrow shows the true answer according to ITSA - whether the judgments should be greater than, less than, or equal to zero. The Empirical Data column in Table 1 shows a simplified summary of the results, which can be compared against the three model predictions.

Nine Datasets

The results for conditions A, C, D, F, G, and I all show that participants' inferences were in line with the predictions of ITSA. In conditions C, D, F, G, and I, all judgments were above zero, as predicted ($ps < 0.001$, BFs > 100). In the flat condition, Condition A, participants appropriately gave causal strength judgments around 0 ($ps > 0.05$, BFs ranged from 1/3 to 1/10). In the flat condition, participants gave positive future use judgements ($ps < 0.001$, BFs > 100) which indicates that the participants thought the patient should stop using the medicine (intervention). Though this numerically differs from causal strength, it makes sense if participants believe that if a medication has no benefit, it should not be used.

Participants' judgments differed from the predictions of ITSA in conditions B, E and H. First, in Condition B in which there is only a pre-intervention slope, the causal strength judgements were higher than zero ($ps < 0.001$, BFs > 100).

Second, in E. Slope Change (Maintain) condition, the causal strength judgements were close to zero ($ps > 0.05$, BFs were between 0.11 to 0.23), and the future use measures were close to zero for the static graph ($p = 0.268$, BF = 0.20) and TbT-number ($p = 0.563$, BF = 0.13). For the TbT-dot there is a bit of evidence of judgments higher than zero; the p-value was significant ($p = 0.019$) but BF was unconvincing at 1.64. The judgments for the dynamic graph were higher than zero ($p = 0.004$, BF = 6.48).

Third, in Condition H where the intercept change is in the opposite direction to the pre-intervention slope, two measures in the static and dynamic graph were close to zero ($ps > 0.05$, BFs between 0.28 and 0.73) and they were lower than zero in two trial-by-trial formats ($ps < 0.001$, BFs > 100).

The rightmost column in Table 1 summarizes the findings qualitatively. None of the theories can explain all of the results, and the patterns are discussed more in the discussion.

Effects of Presentation Formats

Figure 1 shows Bayesian ANOVA results on the top of each graph. For eight out of the nine conditions there were no reliable effects of presentation formats. All but the future use judgments in Condition B and E had a p value larger than 0.05 and a BF between 1 and 1/10 in favor of the null model (the future use judgments in Condition B: $p = 0.023$ but BF = 0.717 in favor of null, in Condition E: $p = 0.015$, BF = 1.13).

We did find a main effect of presentation format in Condition H (causal strength judgment: $p < 0.001$, BF = 60.68; future use judgment: $p < 0.001$, BF = 18.16). The judgements with the static and dynamic graph formats were close to zero but the TbT-dot and TbT-number formats were less than zero. We tested all comparisons by Tukey test and bayesian t test. The static graph group was higher than the TbT-dot group ($p = 0.007$, BF = 17.91) and TbT-number group ($p = 0.038$, BF = 3.79). The dynamic graph group was also higher than the TbT-dot group ($p = 0.003$, BF = 55.53) and TbT-number group ($p = 0.018$, BF = 9.08). There were no differences for the remaining comparisons ($ps > 0.05$, BFs between 1/10 to 1/3). The future use judgments had a similar pattern, except that we didn't find reliable differences in the comparisons dy-

namic graph vs TbT-number (p = 0.284, BF = 0.58) or static graph vs TbT-number (p = 0.082, BF = 2.87).

Discussion

This is the first study to systematically investigate causal learning under interrupted time series data. We manipulated the intercept change, slope change and pre-intervention slope to create various interrupted time series datasets, and presented the data in four formats. Our main finding was that participants were only capable of accurately learning simple interrupted time series where there is no pre-intervention slope or the pre-intervention slope is congruent with the changes caused by the intervention; participants have difficulty controlling for pre-intervention slope when it is incongruent to the changes caused by the intervention. Furthermore, the format only affected learning in one of the datasets.

Comparison of Models There are at least three potential theories for how people assess the influence of intervention. As can be easily seen in our results (Table 1), none of these three theories can explain all the results. It is possible that the participants used a mixture of these reasoning processes of that there are other explanations as well.

The rational way to analyze the causal influence in interrupted time series is to evaluate the changes in slope and intercept from pre- to post-intervention period akin to ITSA, thereby controlling for the pre-intervention slope. An ITSA approach accounts for the results 6 out of 9 conditions and failed to explain the results in Condition B, E and H.

The after-minus-before theory agrees with ITSA and also correctly predicted participants' judgments in five conditions (A, C, D, F, G). Unlike ITSA, it also correctly predicted the results in Condition B. However, it failed to explain the results in E and I. For Condition H, it predicts a small negative finding, so is somewhat in line with the empirical results, but does not explain the differences across the formats.

The post-intervention-trend theory also agrees with ITSA in 5 out of the 9 conditions (A, D, F, G, and I). It correctly predicted the results for 7 conditions. However, this theory cannot explain that participants learned the causal influence in Condition C - Intercept Change, one of the simplest, in which the post-trend is flat. It also explains half of the results in Condition H; it correctly predicts that participants gave negative judgments in the two trial-by-trial conditions, but in the static and dynamic graph conditions participants gave responses around zero.

In sum, none of the three theories can explain all the results on their own, which means that either participants used a combination of these theories, that there are mixtures of different groups of participants, or that there are other theories that better explain the results.

Effects of Incongruency and Format

Aside from Condition B, participants had difficulty correctly assessing causality in the three 'incongruent' conditions. Incongruence is when the the influence of the intervention opposes the direction of the pre-intervention slope.

First, in Condition H (positive datasets), the pre and post trend is negative and the intercept change is positive. However, the static and dynamic graph groups judged the intervention as having no influence on average, and the two trial-by-trial groups inferred a negative causal influence. It is possible that participants in different presentation formats adopted different strategies. In the static and dynamic graph condition, one possibility is that since they could see all the data at once, they could more easily implement the after-minus-before strategy, which in this case revealed only a slight negative influence - close to zero. Another possibility in this condition is that participants were aware both of the positive intercept change (ITSA) and also the negative post-intervention slope, and that they gave judgments near zero because of the opposing thoughts. In the trial-by-trial formats participants tended to give negative judgments. One hypothesis is that they primarily focused on just what was happing 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.

Second, in Condition I, participants tended to correctly infer a positive influence; however, their inferences were quite weak compared to Condition D. In fact, both Conditions I and D had the same degree of slope change according to the regression coefficients, but in Condition I the post-intervention slope was fairly close to flat. This could suggest that participants were using the post-intervention slope rather than the change in slope.

Third, in Condition E the judgments were close to zero but should have been positive. The most reasonable explanation is that participants were biased by the flat post-intervention trend and thought the intervention as ineffective, which fits with the explanation for Condition I.

Conclusions

This research provides a systematic understanding to how people make inferences from interrupted time series results. The participants made a number of errors of reasoning that involve failing to correctly account for a pre-intervention slope and therefore not picking up on the intercept change, slope change, or lack thereof. In one condition we found that format moderated the magnitude of this error. In the future it will be important to study how to improve judgments for these conditions. Furthermore, since none of the theories can independently explain all the results, it will be important to understand if there are subgroups of participants using different approaches, or if there are other theories that can explain more of the findings.

Acknowledgments

This work was supported by NSF BCS 1651330.

References

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.

Derringer, C., & Rottman, B. M. (2016). Temporal causal strength learning with multiple causes. In *Cogsci.*

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.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press.

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.

Soo, K. W., & Rottman, B. M. (2020). Distinguishing causation and correlation: Causal learning from time-series graphs with trends. *Cognition*, *195*, 104079.

White, P. A. (2015). Causal judgements about temporal sequences of events in single individuals. *Quarterly Journal of Experimental Psychology*, *68*(11), 2149–2174.

White, P. A. (2017). Causal judgments about empirical information in an interrupted time series design. *Quarterly Journal of Experimental Psychology*, *70*(1), 18–35.