Cognition 195 (2020) 104079

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Short Communication

Distinguishing causation and correlation: Causal learning from time-series graphs with trends

Kevin W. Soo, Benjamin M. Rottman*

Department of Psychology, University of Pittsburgh, United States

ARTICLE INFO

Keywords:
Causal learning
Time-series graph
Data visualization

ABSTRACT

Time-series graphs are ubiquitous in scientific and popular communications and in mobile health tracking apps. We studied if people can accurately judge whether there is a relation between the two variables in a time-series graph, which is especially challenging if the variables exhibit temporal trends. We found that, for the most part, participants were able to discriminate positive vs. negative relations even when there were strong temporal trends; however, when there is a positive causal relation but opposing temporal trends (one variable increases and the other decreases over time), people have difficulty inferring the positive causal relation. Further, we found that a simple dynamic presentation can ameliorate this challenge. The present finding sheds light on when people can and cannot accurately learn causal relations from time-series data and how to present graphs to aid interpretability.

1. Introduction

Time-series graphs are commonly used in academic journals, popular media, and also for personal decision making such as in fitness and health trackers. When two variables are presented, often the goal is to learn whether and how the two variables are related, for example, tax rates and economic growth (Krugman, 2019) or vaccination and outbreaks (Fig. 1 in Vyse et al., 2002). One question involves determining whether one variable is the cause whereas the other is the effect (Rottman & Keil, 2012; Soo & Rottman, 2014). Another question, which we focus on, is whether there is a positive or negative relation between the two variables (Redelmeier & Tversky, 1996; Soo & Rottman, 2018).

1.1. Time-series data with temporal trends

A major challenge when judging whether two variables are positively or negatively related is that the variables can undergo temporal trends, obscuring the true underlying causal relationship (Yule, 1926). Temporal trends are very common in time-series situations, and controlling for them is critical for personal decision making based on formal or informal "single-subject" research designs (e.g., Sidman, 1960). Consider two examples in which the correlation between X and Y (r_{XY}) leads to incorrect inferences about causation. In the first, a patient takes increasing amounts of medication to cope with increasing pain, which results in a positive correlation even though from day to day an increase in the medication causes a *decrease* in pain. In this example, disease progression is a confound that obscures the negative causal influence. This sort of relationship is represented in the top left causal structure in Fig. 1. Over time (t) the cause (X) and the effect (Y) both increase, but from one day to the next, as X increases, Y decreases, and vice versa.

Consider a second example, the infamous correlation between ice cream sales and drownings. From winter to summer, ice cream sales (X) and drowning deaths (Y) increase, though there is no direct causal relation. The top-right "monotonic trend" in Fig. 1 depicts this sort of causal relation. The temporal trends in both variables can make it falsely appear as if there is a strong causal relationship. However, in a multiple regression predicting Y, X is not significant after controlling for t.

Although we hesitate to say that it is possible to definitively uncover the strength of a causal relation from observational time-series data, we can say that causal structures produce characteristic patterns in time-series data (Fig. 1). A positive causal relation produces "positive transitions"; X and Y tend to increase or decrease together. A negative causal relation produces "negative transition"; X and Y tend to change in opposite directions from one observation to the next. This fact is revealed by two approaches used in time-series analysis to control for temporal trends that are closely related (Shumway & Stoffer, 2011). One approach is to run a regression predicting Y from X, with time as an additional covariate. Another method is to compute change scores (ΔX

* Corresponding author at: Department of Psychology, University of Pittsburgh, LRDC 720, 3939 O'Hara Street, Pittsburgh, PA 15260, United States.
E-mail address: rottman@pitt.edu (B.M. Rottman).

https://doi.org/10.1016/j.cognition.2019.104079
Received 7 June 2019; Received in revised form 12 August 2019; Accepted 18 September 2019
Available online 16 December 2019
0010-0277/

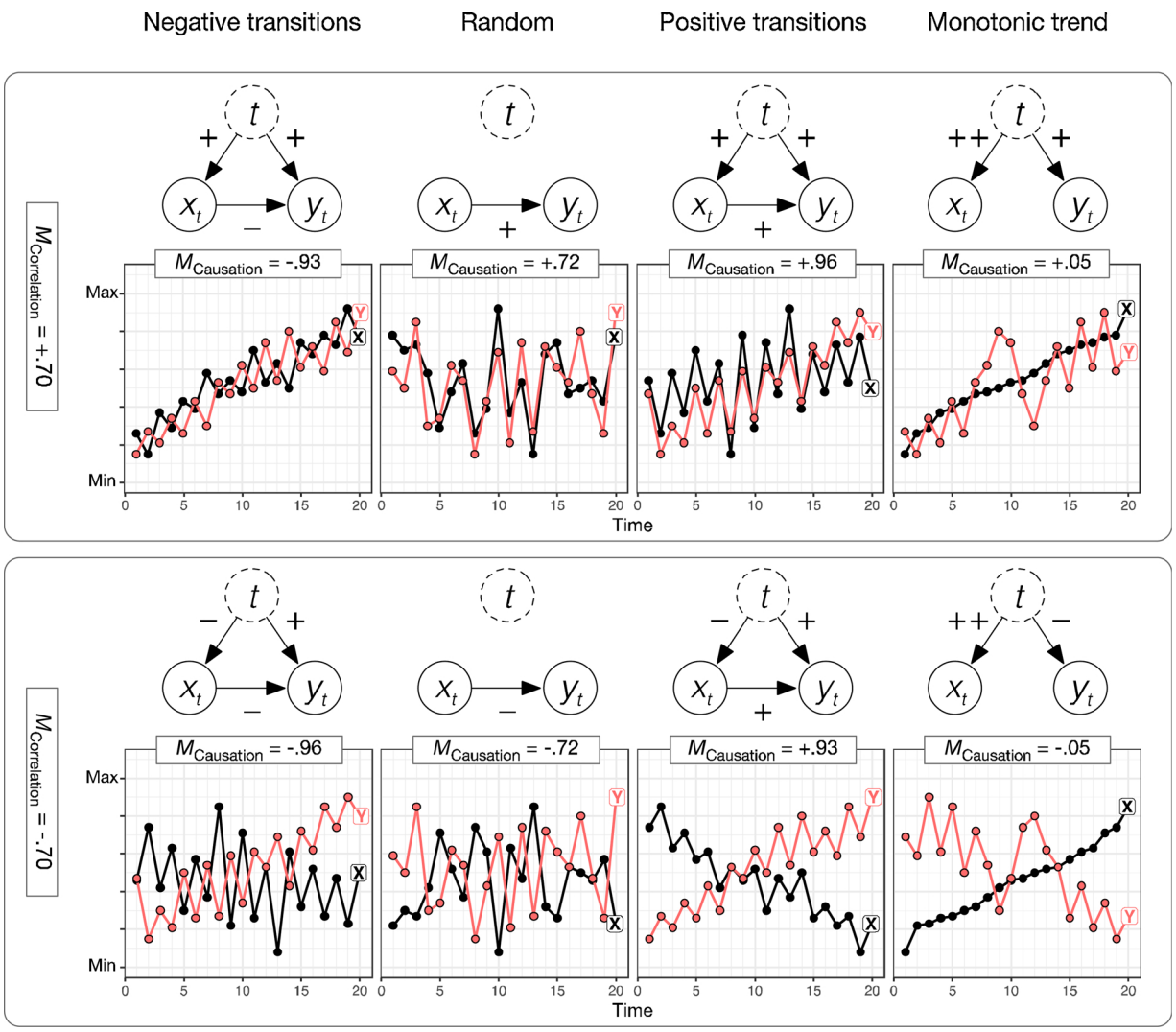

Fig. 1. Causal structures depicting relationships between X, Y, and time *t*, corresponding time-series graphs of example datasets, and the averages of the correlation and causation metrics of stimuli within each condition. In the causal graphs, + (–) means a positive (negative) causal relation and + + means a very strong positive causal relation. The node for time is dashed to represent the fact that it is not literally time that causes X and Y, but this node represents an unobserved variable that causes X and Y, and that also changes over time.

and ΔY), and run a correlation between the change scores ($r_{\Delta X \Delta Y}$)[1] .

1.2. Prior research, motivation, and hypotheses

Soo and Rottman (2018) tested whether people are able to correctly infer whether a causal relation is positive or negative despite strong temporal trends in X and Y. On average, participants were often able to infer the correct relations; their judgments were more strongly influenced by whether the transitions implied positive or negative causation than the correlation between X and Y. However, Soo and Rottman tested this using shapes that changed in size or opacity in a trial-by-trial format, which mimics learning about events in one's own life.

In the current study, we tested whether people can accurately uncover whether a relation is positive or negative from time-series graphs, which is also a critical skill given the prevalence of time-series graphs in science and in the news. We predicted that a static time-series graph would make people attend less to the changes in X and Y and focus more on the overall correlation between X and Y (e.g., noticing that when X is high Y is low, and vice versa, in the negative correlation positive causation example in Fig. 1), making it harder to accurately infer the causal relation between X and Y. However, this hypothesis is not a foregone conclusion; people might naturally scan the graph from left to right, noticing changes in X and Y, possibly enabling accurate inferences.

Furthermore, to understand the mechanism that drives potential differences between formats, we added a hybrid of the trial-by-trial and static graphs, a dynamic time-series graph in which the data are revealed gradually. On the one hand, if participants in the static graph condition already scan the graph from left to right and notice the changes at each trial, then presumably in both of the graph conditions they will perform similarly to the trial-by-trial condition. On the other hand, if participants just notice the overall correlation in the static graph condition, they might perform similarly in the dynamic graph condition (the fully revealed dynamic graph looks identical to the static graph), or the sequential revealing of the data in the dynamic graph condition might help people notice the changes in X and Y on each trial in which case they would perform similar to the trial-by-trial condition.

There is one other important difference between this study and our prior research (Soo & Rottman, 2018). We added a fourth *monotonic*

[1] In Soo and Rottman (2018), we called this strategy $r_{\Delta Continuous}$, and contrasted it against another very similar strategy called $r_{\Delta Binary}$ in which the changes were simply coded as +1 for increase and -1 for decrease. For this paper, we did not compare the two, so we just call it $r_{\Delta X \Delta Y}$.

Table 1
Mean and (SD) of Correlation and Causation Stimuli Properties.

Condition (causation)	Degree of causation	Condition (correlation)			
		Positive correlation r_{XY} = .70 (0.01)		Negative correlation r_{XY} = -.70 (0.01)	
		$r_{\Delta X \Delta Y}$	Regression	$r_{\Delta X \Delta Y}$	Regression
Negative transitions	Strongly negative	−.93 (0.03)	−.79 (0.18)	−.96 (0.02)	−.75 (0.13)
Random Order	Depends on correlation	.72 (0.08)	.70 (0.03)	−.72 (0.08)	−.70 (0.03)
Positive transitions	Strongly positive	.96 (0.02)	.75 (0.13)	.93 (0.03)	.79 (0.18)
Monotonic trend in X	Close to zero	−.02 (0.19)	.04 (0.87)*	.02 (0.19)	−.04 (0.87)*
Monotonic trend in Y	Close to zero	.13 (0.29)	.01 (0.03)	−.13 (0.29)	−.01 (0.03)

Note. For Regression, the reported numbers are the standardized regression weights for X on Y after controlling for time. *The high standard deviation for these two cases is due to the fact that when there is a monotonic trend in X, X and *t* are highly correlated. The multicollinearity can cause very large positive or negative standardized beta weights even though the effects are not significant. Six of the 20 datasets had very large standardized regression weights.

trend condition (Fig. 1), in which one of the variables increases or decreases smoothly over time. The monotonic trend condition tests whether people can infer the *absence* of a causal relation, which has been challenging for people with other sorts of time-series data (Redelmeier & Tversky, 1996).

2. Method

2.1. Participants

Participants were recruited on Amazon Mechanical Turk in two batches of 151 and 152.[2] The experiment lasted 7–10 minutes and participants were paid $1.40.

2.2. Cover story

Participants judged how the dosage of a drug (X) influenced the size of a microorganism (Y). Each dataset consisted of 20 observations, framed as 20 sequential days. After viewing the entire dataset, participants estimated the causal strength of the drug on a scale from 8 ("high levels of the drug strongly cause the microorganism to increase in size") to -8 ("high levels of the drug strongly cause the microorganism to decrease in size"), with zero indicating there was no causal relationship. Each scenario was framed as involving a different drug and microorganism.

2.3. Design

The experiment used a 3 (presentation format: trial-by-trial, dynamic graph, or static graph; between-subjects) × 2 (correlation: positive vs. negative; within-subjects) × 5 (causation: negative transitions, random order, positive transitions, monotonic trend in X, or monotonic trend in Y; within-subjects) design. We ran two batches of participants because we only realized after the first that we could create the monotonic trend conditions. We report the study with the data aggregated for concision.

2.4. Stimuli

2.4.1. Manipulation of correlation and causation

We created time-series data that implied different degrees of correlation and causation (see Fig. 1 for examples). We first created 20 datasets, and then manipulated the order of the observations for the different conditions. Thus, while at a high level we manipulated the causation (positive or negative) implied by the dataset, we accomplished this by manipulating the order of the observations.

All datasets consisted of 20 observations of a cause (X) and effect (Y), each of which could take on values between 0 and 100. We generated 20 datasets with r_{XY} = .70 ± .01 using the R package ecodist. For the negative correlation (r_{XY} = -.70) condition, we flipped the values of X around the midpoint.

We manipulated the causation implied by the datasets by varying the order of the observations. Table 1 shows the mean and standard deviation of the measures of correlation and causation for the datasets across all 10 conditions. Table 1 reports two ways of trying to infer the strength of the causal relation by controlling for temporal trends: $r_{\Delta X \Delta Y}$ and regression with time as a linear covariate (Shumway & Stoffer, 2011). These two are closely related. For the analyses, we used $r_{\Delta X \Delta Y}$ for consistency with prior research (Soo & Rottman, 2018) and because some of the conditions have high standard deviations in the regression coefficients due to multicollinearity. Using regression does not meaningfully change the results.

In the random condition, the 20 trials were in a random order, resulting in datasets in which the measures of correlation and causation are related. In the positive correlation random order condition, usually when X increased Y increased, and when X decreased Y decreased (positive transitions). In the negative correlation random order condition, most of the transitions were negative.

In the positive transitions condition, the trials were ordered so that increases in X were always accompanied by increases in Y, resulting in extremely strong positive causation. In the negative transitions condition, the trials were ordered so that X and Y always changed in opposite directions, resulting in extremely strong negative causation.

Finally, there were two monotonic trend conditions; either X or Y always increased or decreased across the 20 trials.[3] When creating the datasets, those that that had repeated values of X or Y (e.g., X was exactly 56 on two trials) were slightly modified (e.g., the value of X on one trial was changed to 58) so that the trials could be ordered monotonically. This change in the dataset was made in all the conditions. According to both metrics of causation (Table 1), these datasets imply minimal if any causal influence of X on Y.

The trials in each dataset were presented forwards or in reverse (i.e. from 1 to 20 or 20-1) randomly for each scenario. For example, in the negative-transitions positive-correlation condition in Fig. 1, both variables increase over time. Reversing the order means that both variables decrease over time, but the stimuli still consisted of negative transitions and positive correlation.

2.4.2. Manipulation of presentation format

In the *trial-by-trial* condition, states of the cause and effect were mapped to gauges (Fig. 2A), similar to Experiments 2 and 3 from Soo

[2] We recruited 150 but a few participants completed the experiment without claiming payment.

[3] The monotonic trend in Y condition looks very similar to the monotonic trend in X condition in Fig. 1 but is not depicted for brevity.

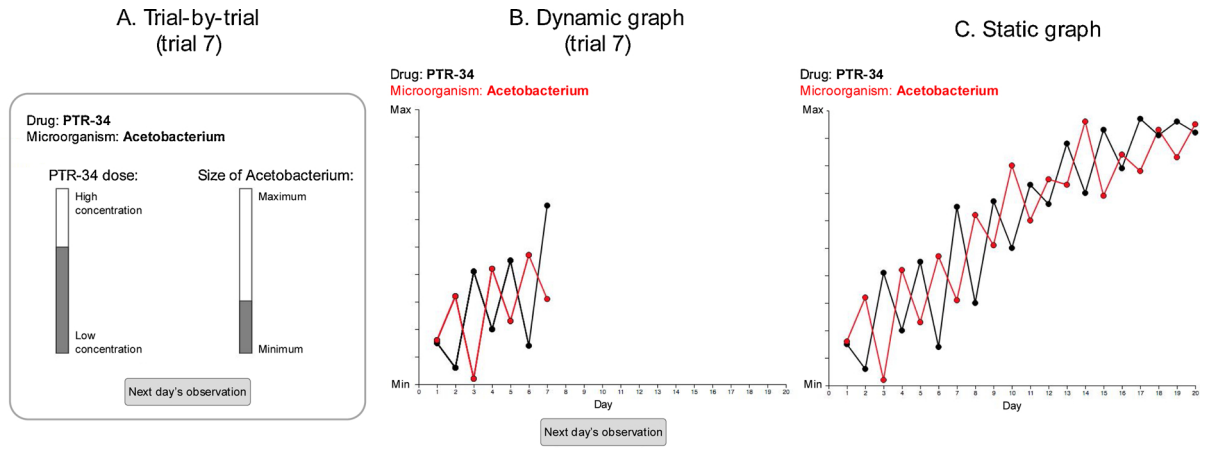

Fig. 2. Stimuli shown to participants in different presentation formats. In (A) and (B), observations are revealed sequentially after participants clicked "Next day's observation". In (C), participants view the entire graph for 40 s.

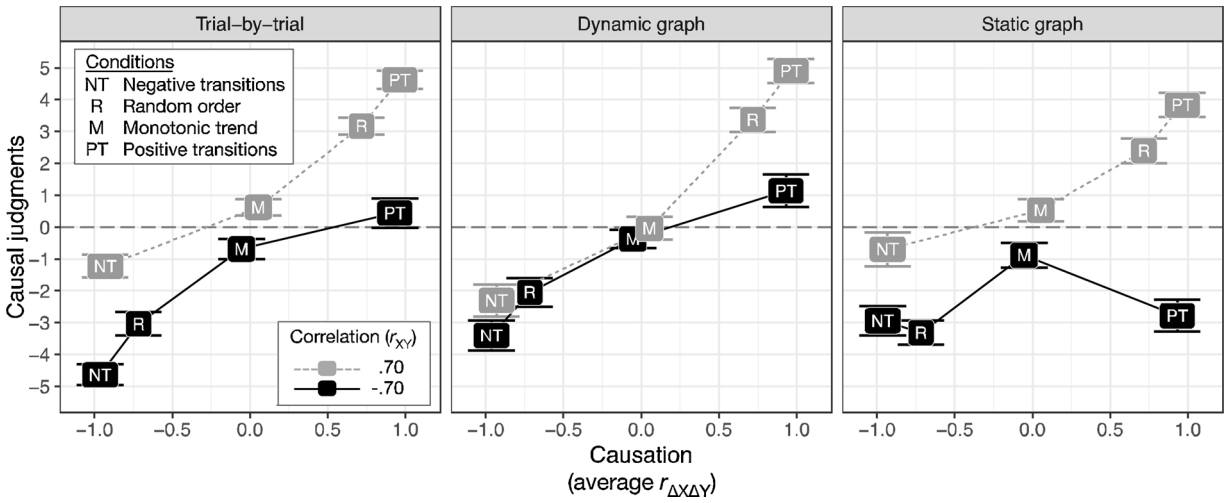

Fig. 3. Means of participants' causal strength judgments by condition. Error bars represent standard errors.

and Rottman (2018)b). Participants clicked a button to reveal the observation for the next "day", at which point the two gauges would slide up or down to display the data for the next observation. Participants waited at least two seconds before advancing to the next trial.

The *static graph* condition (Fig. 2C) presented the data in a line graph. The graph was kept on the screen for 40 s, at which point the graph disappeared and participants made their causal judgment.

The *dynamic graph* condition (Fig. 2B), was identical to the *static graph* condition, except that the observations were revealed sequentially from left to right. Fig. 2B shows what a participant might see on trial seven. Participants clicked a button to reveal each new observation, and waited at least two seconds before clicking again.

2.4.3. Attention verification task

The trial-by-trial and dynamic graph conditions required participants to watch the screen and click to advance to the next observation. In contrast, it was possible that participants in the static graph condition to look away from the screen during the 40 s because they do not need to click to advance. To prevent this, we included an attention verification task in all three conditions. In the trial-by-trial and dynamic graph conditions, somewhere between trials 15–18, a five-letter word appeared for three seconds. Participants had to report the word prior to making their causal judgment. In the static graph condition, the word occurred between 25–35 seconds after the start of the scenario.

3. Results

Stimuli, data, and analysis scripts are available at https://osf.io/gyvdw/.

Participants correctly recalled the word at high rates in the trial-by-trial (97%), dynamic graph (96%), and static graph (94%) conditions. Thus, we did not exclude any of the data in our analyses.

The means of participants' causal judgments are displayed in Fig. 3. The conditions have been positioned along the x-axis according to the average level of causation (Table 1). Perfect performance – judging causal strength entirely based on $r_{\Delta X \Delta Y}$ – would appear as a diagonal line. In Fig. 3, we collapsed the two monotonic trend conditions within the positive correlation condition and the two within the negative correlation condition because they are highly similar in causation.

In the analyses, we used the actual degree of correlation and causation (measured with r_{XY} and $r_{\Delta X \Delta Y}$, respectively) for each individual dataset as predictors. Effect sizes are presented with R^2_{NSJ} for random-effects models (Jaeger, Edwards, Das, & Sen, 2017) and are also converted into Cohen's *d*.

We first ran an overall regression predicting participants' causal judgments from each dataset's correlation and causation values, the presentation format, and the interactions, with by-participant random slopes for correlation and causation and their interaction for repeated measures. Most important for our predictions, there was a significant interaction between causation and presentation format, $F(2,$

Table 2
Regressions of subjects' causal strength ratings for the three presentation format conditions.

Presentation format	Correlation (r_{XY})			Causation ($r_{\Delta X \Delta Y}$)			Interaction ($r_{XY} * r_{\Delta X \Delta Y}$)			Whole Model
	b	R^2_{NSJ}	*d*	*b*	R^2_{NSJ}	*d*	*b*	R^2_{NSJ}	*d*	R^2_{NSJ}
Trial-by-trial	1.96***	.13	0.78	2.72***	.26	1.18	0.30	.002	0.09	.40
Dynamic graph	1.10***	.03	0.35	2.94***	.20	1.01	1.02***	.02	0.25	.27
Static graph	2.54***	.13	0.79	1.27***	.04	0.43	1.32***	.02	0.31	.22

Note. ***$p < .001$ level. Confidence intervals are reported at the 95% level.

300.95) = 11.02, $p < .001$[4], and there were also other two and three-way interactions.

Because of the two and three-way interactions, for clarity and concision we skip next to reporting regressions for each of the three formats (Table 2). Each regression included by-subject random intercepts and slopes for correlation, causation, and their interaction. The effect of correlation is significant in all three conditions, but smallest in the dynamic graph condition. As expected, the effect of causation is significant in all three conditions, but smallest in the static graph condition, which implies that participants performed worst in this condition. In addition, there was an interaction between correlation and causation in in the dynamic and static graph conditions, but not in the trial-by-trial condition.

Finally, we statistically compared the three presentation formats. First, the influence of correlation was weaker in the dynamic graph format compared to both the trial-by-trial (b = -0.86 [0.22, 1.50], p = .01, R^2_{NSJ} = .01, d = 0.16), and static graph formats (b = -1.42 [0.65, 2.18], p < .001, R^2_{NSJ} = .01, d = 0.22). There was no difference in the influence of correlation between the trial-by-trial and static graph formats (p =.13).

Second, the primary hypothesis was that presentation format would moderate the influence of causation. As predicted, participants' judgments in the static graph format were less accurate (less sensitive to causation) than in the trial-by-trial format (b = -1.45 [-2.15, -0.76], p < .001, R^2_{NSJ} = .02, d = 0.28) and the dynamic graph format (b = -1.67 [-2.49, -0.84], p < .001, R^2_{NSJ} = .02, d = 0.29). There was no difference between the trial-by-trial and dynamic graph formats (p =.57).

Third, we tested whether presentation format moderates the interaction between correlation and causation. On inspection of Fig. 3, in the trial-by-trial condition, the lines representing the effect of causation in the negative and positive correlation conditions are parallel meaning that the influence of causation is the same whether there is a positive correlation or a negative correlation. However, in the static and dynamic graph conditions, the influence of causation is weaker (flatter) in the negative correlation condition. In fact, the interaction between correlation and causation was weaker in the trial-by-trial compared to the dynamic graph format (b = -0.74 [-1.38, -0.10], p = .025, R^2_{NSJ} = .002, d = 0.09), and the static graph format (b = -1.03 [-1.71, -0.35], p = .003, R^2_{NSJ} = .005, d = 0.14). There was no difference between the dynamic and static graph formats (p = .45).

This finding appears to be largely driven by the positive-transition negative-correlation condition, especially in the static graph format. In this condition, participants actually inferred a negative causal relationship even though all the transitions were positive. The positive-transition negative-correlation condition was lower in the static graph than the dynamic graph condition, $t(201) = 5.44$, $p < .001$, d = .76, and the trial-by-trial condition, $t(199) = 4.72$, $p < .001$, d = .67, and there was no difference between the dynamic graph and trial-by-trial conditions, p = .31.

We suspect that the reason why the judgments were so inaccurately low in the positive-transition negative-correlation static graph condition was that the trends in the variables go in opposite directions so the two lines overlap only briefly (see Fig. 1). This brief overlap makes it hard to notice that the cause and the effect change in the same direction from one trial to the next (positive transitions). Instead, the overwhelming impression of the graph is that the variables trend in opposite directions.

This is an interesting finding because it is not that people are always bad when correlation and causation conflict in static graphs. When the trends move in the same direction (e.g., in the positive-correlation but negative-transition condition), it is fairly easy to notice how the cause and effect move in opposite directions, and indeed on average participants inferred a slightly negative causal relation.

4. General discussion

Given how much time-series data is collected not only for policy decisions (e.g., the economy or business) but also for personal decisions (e.g., health tracking smartphone apps), it is critical to understand if people are able to draw accurate causal inferences, and how to help them do so. Prior research has investigated how to improve interpretability of time series graphs (Javed, McDonnel, & Elmqvist, 2010; Wang, Han, Zhu, Deussen, & Chen, 2018), though has not focused on how people infer *causal relations* from time series graphs.

We found that people are fairly accurate at inferring causal relations, which is welcome news. However, we found one situation in which people have considerably difficulty – when one variable exhibits an increasing trend, the other a decreasing trend, and there is a positive causal relation. This is likely because the lines barely overlap, making the relations between short-scale transitions in the two lines hard to notice. When the time-series graph was dynamically revealed, participants performed quite well even in the problematic case. This dynamic presentation could be easily used in electronic media (e.g., television, internet).

4.1. Caveats

In this paper we have used strong language about how people should infer causal strength in time series data despite the fact that making inferences from time series data is notoriously fraught. Here we briefly discuss some nuances and caveats to these claims. First, this paper is not about inferring the causal structure or direction of causality; it is only about inferring the strength of the relation between two observed variables. Standard time series data analysis requires controlling for temporal trends when assessing the relation between two variables, lest the correlation due to the temporal trends overrides the direct causal relations. The point of the paper is to uncover which formats of presentation aid people in controlling for the temporal trends.

Second, there are a number of ways in which the inference task for real world data can be more complicated than in the current task. If the causal relation between X and Y has a delay, then the learner must "shift" the data to calculate the relation within the right window of time. People

[4] This key finding was significant in both the first $F(2, 148.37) = 10.23$, $p < .001$ and second batches of participants, $F(2, 148.62) = 4.78$, p = .010. Thus, we decided analyze both batches together. We do not report effect sizes for this analysis because we do not know of any effect size measures for random effect models that handle factors. For the entire model, R^2_{NSJ} = .24.

generally know to perform this shift (Buehner & May, 2002; Hagmayer & Waldmann, 2002), though it is an open question as to whether shifting may make it harder to notice short-term changes in X and Y. We contend that these issues are orthogonal; if a learner believes that there is a delay, then they should account for the delay in addition to controlling for the temporal trends. (It is trivial to add a shift into the $r_{\Delta X \Delta Y}$ strategy; just calculate the correlation between the change in X at one point of time and the change in Y at a later point in time.)

It is also possible that X could cause Y, and Y could cause X, or both, or there could be a third variable that causes both of them. Again, the goal for this paper was not to study how people infer causal structure, and we believe that accurately assessing the strength of the relation in any of these situations requires controlling for temporal trends in addition to accounting any delay(s) in the relation(s).

In summary, although inferring the strength of the relation is likely to be harder in real world datasets than in the current datasets, the current results suggest that in many situations people do have the capacity to control for temporal trends, and dynamic presentations can help them do so.

References

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning, 8*(4), 269–295.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30*(7), 1128–1137.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics, 44*(6), 1086–1105.

Javed, W., McDonnel, B., & Elmqvist, N. (2010). Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 927–934. https://doi.org/10.1109/TVCG.2010.162.

Krugman, P. (2019). *The economics of soaking the rich.* January 5. The New York Times Retrieved from https://www.nytimes.com/2019/01/05/opinion/alexandria-ocasio-cortez-tax-policy-dance.html.

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

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology, 64*(1–2), 93–125. https://doi.org/10.1016/j.cogpsych.2011.10.003.

Shumway, R. H., & Stoffer, D. S. (2011). *Time series analysis and its applications.* New York: Springer.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York: Basic Books.

Soo, K. W., & Rottman, B. M. (2014). Learning causal direction from transitions with continuous and noisy variables. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.). *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1485–1490). Austin, TX: Cognitive Science Society.

Soo, K. W., & Rottman, B. M. (2018). Causal strength induction from time series data. *Journal of Experimental Psychology General, 147*(4), 485–513.

Vyse, A. J., Gay, N. J., White, J. M., Ramsay, M. E., Brown, D. W. G., Cohen, B. J., ... Miller, E. (2002). Evolution of surveillance of measles, mumps, and rubella in England and Wales: Providing the platform for evidence-based vaccination policy. *Epidemiologic Reviews, 24*(2), 125–136. https://doi.org/10.1093/epirev/mxf002.

Wang, Y., Han, F., Zhu, L., Deussen, O., & Chen, B. (2018). Line graph or scatter plot? Automatic selection of methods for visualizing trends in time series. *IEEE Transactions on Visualization and Computer Graphics, 24*(2), 1141–1154. https://doi.org/10.1109/TVCG.2017.2653106.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society, 89*(1), 1–63.