Comparing Mediation Inferences and Explaining Away Inferences on Three Variable Causal Structures

Cory J. Derringer (cjd78@pitt.edu)
Benjamin M. Rottman (rottman@pitt.edu)
Department of Psychology, University of Pittsburgh
3939 O'Hara Street, Pittsburgh, PA 15260 USA

Abstract

People reliably make two errors when making inferences about three-variable causal structures: they violate what is known as the Markov assumption (mediation) on causal chains and common cause structures, and fail to sufficiently 'explain away' on common effect structures. Our goal for the present study was to quantitatively compare these two errors after subjects have learned the statistical relations between three variables using procedures designed to maximize the accuracy of their learning and inferences. Aligning with prior research, we found that subjects violated the Markov assumption, and did not sufficiently explain away. We also found judgments about mediation were worse than judgments about explaining away for one inference, but better for another, suggesting that people are not uniquely worse at reasoning about one structure than another. We discuss the results in terms of a theory of cue consistency.

Keywords: causal learning; causality; causal structure; Markov assumption; explaining away

Introduction

Causal learning is a ubiquitous part of our everyday lives, from determining the efficacy of a new medicine to figuring out if our new diet/fitness routine is working. Further, often it is important to understand the causal structure and statistical relations among several variables in order to intervene effectively; if one knows that exercise only leads to weight loss because it creates a caloric deficit, one knows to avoid other activities that might negate that pathway (e.g., eating an extra helping of food). In the current research we examined how well people understand the statistical relations among different three-variable structures. Specifically, we compared the accuracy of judgments of $P(X_2|Y, X_1)$ on mediation and common effect structures (Figure 1).

Understanding how well people make this inference provides insight into how well people understand the differences in the multivariate distribution (the ways that all three variables are statistically related to each other) for different sorts of causal structures. In particular, the goal for this research is to understand if people have a specific difficulty understanding the multivariate distribution for some causal structures (potentially common effect structures) but are better for others (potentially mediation structures).

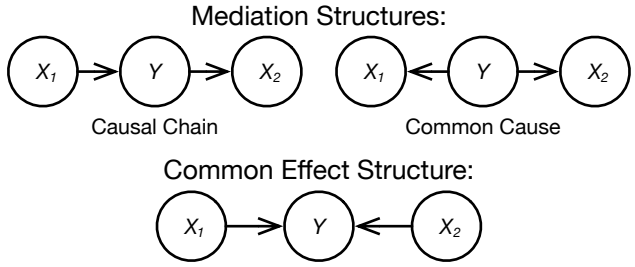

Figure 1: Examples of mediation (causal chain, common cause) and common effect causal structures.

The Markov Assumption on Chain and Common Cause Structures

An important feature of causal chain and common cause structures is that the middle variable Y blocks off any direct connection between X_1 and X_2. This feature, called the Markov assumption, means that once a learner knows the state of Y, X_1 is no longer predictive of X_2. For example, consider a chain in which exercise (X_1) causes a caloric balance (whether caloric intake is greater or less than caloric expenditure) (Y), which influences body weight (X_2). The probability of losing weight given that caloric intake is less than expenditure is high, regardless of exercise (because exercise only affects weight via caloric balance).

People reliably violate the Markov assumption; they act as if X_1 and X_2 have a hidden connection beyond Y in a chain or common cause structure (e.g., Park & Sloman, 2013; Rehder, 2014; Rehder & Burnett, 2005; Rottman & Hastie, 2014; 2016). The classic violation involves estimating that $P(X_2=1|Y=1, X_1=1) > P(X_2=1|Y=1, X_1=0)$, when in reality they are equivalent.[1] We call the chain and common cause 'mediation structures' as they feature the same statistical relations between the variables (Steyvers et al., 2003).

However, the violation of the Markov assumption, though reliable, is not always large. Rehder and Waldmann (2017) found mean differences between the two judgments of .06 and .15 in two different conditions. Rottman and Hastie (2016) found differences of .19 and .13 in two studies. Park and Sloman (2013; Experiment 3) found some judgments that did not show violations (-01, .01) and some that did (.19, .21).

[1] $P(X_2=1|Y=1, X_1=1)$ can be read as 'the probability that $X_2=1$ knowing that $Y=1$ and $X_1=1$.

Explaining Away on a Common Effect Structure

Common effect structures (e.g., Figure 1) differ from mediation structures in that it is no longer true that $P(X_2=1|Y=1,X_1=1)=P(X_2=1|Y=1, X_1=0)$; instead, $P(X_2=1|Y=1, X_1=1) < P(X_2=1|Y=1, X_1=0)$. For example, assuming that high exercise ($X_1=1$) and low caloric intake ($X_2=1$) both cause weight loss ($Y=1$), the normative relationship between exercise and caloric intake hinges on whether weight loss is known. If the state of Y is unknown, X_1 and X_2 are uncorrelated. However, because weight loss can be caused by either exercise or dieting, if $Y=1$ (a person is losing weight), then if they were not dieting, they probably were exercising, or vice versa: $P(X_2=1|Y=1, X_1=1) < P(X_2=1|Y=1, X_1=0)$.

Previous research has found that people have difficulty making explaining away judgments. Either the amount is insufficient – the judgments of $P(X_2=1|Y=1, X_1=1)$ are lower than $P(X_2=1|Y=1, X_1=0)$, but not sufficiently lower – or the two judgments are not statistically different from each other on average, or sometimes the judgments of $P(X_2=1|Y=1, X_1=1)$ are *higher* than $P(X_2=1|Y=1, X_1=0)$. For example, in Rottman and Hastie's (2016) Experiment 1, the normative difference – $P(X_2=1|Y=1, X_1=0)$ - $P(X_2=1|Y=1, X_1=1)$ – was .40, but there was no significant explaining away on average; the mean amount was -.01. In their Experiment 2, the normative amount was .67, and the mean amount of explaining away was significant but was only .17. Rehder and Waldmann (2017) also found insufficient explaining away; the degree of explaining away was about .15 and .25 in two conditions, whereas it should have been .46.

Cue Consistency Theories

One explanation for both the violations of the Markov assumption and poor explaining away is a theory we call "cue consistency", which suggests when more of the known cues are in state 1, the learner is more likely to infer that the unknown cue is 1, and when more of the known cues are in state 0, the learner is more likely to infer that the unknown cue is 0. Rehder (2014) called this an "associative bias." Rottman and Hastie (2016) called this the "monotonicity assumption," and Rehder and Waldmann (2017) called it the "rich-get-richer principle." We think that all of these principles are essentially the same and we are using "cue consistency" to capture all of these meanings.

Cue consistency can be thought of as systematically misunderstanding the multivariate structure among the three variables as being more similar to the bivariate structure among pairs of variables than it really is. This cue consistency principle explains why people tend to judge that $P(X_2=1|Y=1,X_1=1) > P(X_2=1|Y=1, X_1=0)$ for the mediation structures even though they should be equivalent. The logic is that they think that X_1 provides information about X_2 like it does when judging the bivariate relation between the two.

It also explains why people tend to insufficiently judge that $P(X_2=1|Y=1, X_1=1) < P(X_2=1|Y=1, X_1=0)$ for the common effect structure. People tend to overestimate $P(X_2=1|Y=1, X_1=1)$ or underestimate $P(X_2=1|Y=1, X_1=0)$ or both. The logic is that even though X_1 and X_2 and unrelated in the bivariate relation, they are *negatively* related in the multivariate distribution. Cue consistency theories cannot explain all the known biases in judgments on three variables (Rottman & Hastie, 2016; Table 11), but are plausible explanations for some of the most studied biases including these.

Open Questions

Directly comparing the structures. Comparing the two prior sections on violations of the Markov assumption and insufficient explaining away, the violations of the Markov assumption are on the order of 0-.21; however some of the explaining away judgments were .30 or .40 less than they should have been, and there is other evidence of explaining away judgments *significantly in the wrong direction* (Rottman & Hastie, 2014). This raises the question of whether people understand common effect structures worse than mediation structures. If so, this finding would be important because it could be used to predict when people are especially likely to make poor judgments. To test this, we compared $P(X_2|Y, X_1)$ judgments on the two structures.

This could not be accomplished in previous studies because the normative judgments for the two structures were not equated. In Rottman and Hastie's (2016) Experiment 1b, $P(X_2=1|Y=1, X_1=1)$ was .75 for the mediation structures but .60 for the common effect so they could not be compared, and $P(X_2=1|Y=1, X_1=0)$ was .75 for the mediation structures but 1.00 for the common effect. Potentially some of the differences in these comparisons could be due to the fact that subjective experiences of probability do not map linearly onto objective probabilities (Tversky & Kahneman, 1992).

In the current study, we created two different versions of the common effect structure and one version of the mediation structure, so all of the normative judgments needed for comparison are .80 to .83. In order to accomplish this, one of the inferences actually had to be .17, which we flipped to the upper half of the scale (.83) to compare with other judgments.

Is cue consistency a bias of learning or judgment? Another important question is whether the biased inferences are merely due to a low-level bias at the time of judgment, or if this bias might actually arise through learning? We will lay out three (*not* mutually exclusive) possibilities of how such a cue consistency bias could play out.

In some studies (e.g., Park & Sloman, 2013; Rottman & Hastie, 2016), all three variables had the same two possible states (e.g., all three variables could be high or low, which we represent as 1 vs. 0). In these studies, a reasoner could do a very simple and low level of perceptual matching: if $Y=1$ and $X_1=1$, then probably $X_2=1$ as well.

In other studies, the states of the three variables are counterbalanced, such that the cue consistency bias can't involve a simple perceptual matching. For example, Rehder and Waldmann (2017; also Rehder, 2014) used the following variables: high or low interest rates, small or large trade deficits, and high or low retirement savings. Furthermore, the stimuli were counterbalanced such that some subjects were told that low interest rates cause high retirement savings, others were told that high interest rates cause high retirement

savings etc. Suppose that a subject was told that *high* interest rates cause *large* trade deficits, which cause *low* retirement savings. In this case, the typical violation of the Markov assumption would be inferring that *P*(*high* interest rates|*large* trade deficits, *low* retirement savings) > *P*(*high* interest rates|*large* trade deficits, *high* retirement savings). The violation goes against the simple perceptual matching, but instead follows which states are causally related. We still call this a cue consistency effect, but consistency refers to the believed causal relations rather than perceptual matching.

When a study uses this counterbalancing, there are two ways that a bias could arise. First, the bias could arise if the instructions in the experiment tell subjects the causal relations (e.g., telling them that *high* interest rates cause *large* trade deficits, which cause *low* retirement savings). Second, even if subjects are not told the causal relations, the bias could arise if subjects learn the statistical relations from experience (e.g., *high* interest rates are correlated with *large* trade deficits, which are correlated with *low* retirement savings). Rehder and Waldmann (2017) compared conditions in which participants were told the relations and/or learned the relations from data; the inferences were most accurate (smallest violations of the Markov assumption, strongest explaining away) in the data-only condition suggesting that the bias might primarily arise from instruction.

In the current study, we wanted to maximize the possibility that subjects would give accurate judgments, and minimize the possibility that they are biased by cue consistency. For this reason, subjects learned the statistical relations between the variables from experience, the states of the variables were counterbalanced so that there cannot be an overall perceptual effect of cue consistency, and subjects were not told the relations between the variables in instructions. This means that if they exhibit violations of the Markov assumption, it must arise from the learning process, and if they insufficiently explain away, this must have to do with incorrect or insufficient learning, not a simple perceptual bias or bias from the instructions. By setting up a situation to maximize the accuracy of judgments, we have a fairly pure comparison of mediation vs. common effect inferences.

Method

Participants

Participants (n=230) were recruited via MTurk and were paid $3.00 for their participation. They were also paid bonuses for accurate judgments. All participants were located in the United States, had previously completed at least 100 tasks on MTurk, and had a task approval rate of at least 95%. The study took approximately 20 minutes to complete.

Stimuli and Design

There were three sets of learning data: a mediation structure and two common effect structures (Table 1). These three datasets were chosen such that the key inferences (bold in Table 1) were all normatively between .80 - .83, so that they could be compared. In order to find datasets with these key inferences, we targeted certain parameters that we knew would produce inferences in this range. With a limited number of trials, the datasets cannot fit the parameters exactly. In Table 1, we listed the number of trials presented to participants, and the ideal number of trials if the parameters were followed exactly, to show the closeness of fit.

The data for the mediation structure fit equally well with the chain or common cause structure, which is why we call it the more generic 'mediation' structure. The generative parameters for a common cause are $P(Y=1) =. 5$, $P(X_{1,2} = 1 \mid Y = 1) = .85$, $P(X_{1,2} = 1 \mid Y = 0) = .15$, or for a chain, $P(X_1 = 1) = .5$, $P(Y = 1 \mid X_1 = 1) = P(X_2 = 1 \mid Y = 1) = .85$, and $P(Y = 1 \mid X_1 = 0) = P(X_2 = 1 \mid Y = 0) =.15$.

The idealized Common Effect 1 parameters are $P(X_1=1)=P(X_2=1)=.40$, Power-PC causal strengths of .68 (Cheng, 1997), with an unobserved background cause of probability .10. The idealized Common Effect 2 parameters are: $P(X_1=1)=P(X_2=1)=.13$, Power-PC causal strengths of .73, with an unobserved background cause of probability .06. In order to fit the parameters closely, the datasets had slightly different numbers of trials ranging from 55-58 (Table 1).

Table 1: Learning Data and Key Inferences

X_1	Y	X_2	Mediation Structure	Common Effect 1	Common Effect 2
Actual and (ideal) number of trials in learning data					
1	1	1	20 (19.9)	8 (8.0)	1 (0.8)
1	1	0	4 (3.5)	9 (9.2)	5 (4.5)
1	0	1	1 (0.6)	1 (0.8)	0 (0.1)
1	0	0	4 (3.5)	4 (4.0)	2 (1.5)
0	1	1	4 (3.5)	9 (9.2)	5 (4.5)
0	1	0	1 (0.6)	2 (1.8)	3 (2.6)
0	0	1	4 (3.5)	4 (4.0)	2 (1.5)
0	0	0	20 (19.9)	18 (18.0)	39 (39.4)
Total:			58 (55)	55 (55)	57 (55)
Key Inferences					
$P(X_2=1\|Y=1, X_1=1)$			**.83**	.47	.17→**.83**
$P(X_2=1\|Y=1, X_1=0)$			**.80**	**.82**	.63

Note. → means 'recoded as.' Bold highlight the key inferences in the range of .80 - .83. Ideal number of trials if the parameters were followed exactly in parentheses.

The key inferences were calculated directly from the actual data subjects saw, not the idealized parameters.

Each trial in the datasets described a fictional microbe. The microbes could have one of several sets of features (e.g., cytoplasm color, long or short cilia, and a circle or oval shaped nucleus). Overall, there were nine different types of variables grouped into three clusters of three, and two of the three clusters were randomly assigned to the datasets, and the three variable labels were randomly assigned to X_1, X_2, and Y.

The matching of microbe features to variables in the causal structure was determined randomly for each participant. For

example, one participant might see a dataset in which cytoplasm color corresponds to X_1, whereas for another participant that feature could correspond to Y in the underlying causal structure. Furthermore, the mapping of the color of the cytoplasm (blue vs. red) to 1 vs. 0 in Table 1 was random. Collectively, this randomization means that participants would not have a perceptual reason to infer a particular value for X_1 given the perceptual states of Y and X_2.

Procedure

Participants were told that they would be learning about the features of microbes (e.g., cilia length, cytoplasm color, nucleus shape). Their task was to figure out how the properties related to each other; they were not told if or how they were related. Participants learned about two datasets. (In reality, there was a fourth dataset designed for a question not analyzed here; all subjects learned about two of the four.) Each scenario (dataset) involved two phases.

In the learning phase, participants engaged in trial-by-trial learning of the three features. Subjects made predictions of each of the three features on each trial, which was intended to make the task engaging and to encourage learning. During each trial, participants were first shown a blank template of a microbe (a grey circle). They made a probability judgment about a target feature (Figure 2A). They were then shown the state of another feature and asked about the target feature again (Figure 2B). Finally, they were shown information about the third feature and asked about the target feature again (Figure 2C), and were given subsequent feedback about its state (Figure 2D). Though these judgments give us detailed measures of the progression of learning, we do not analyze these judgments in the current paper.

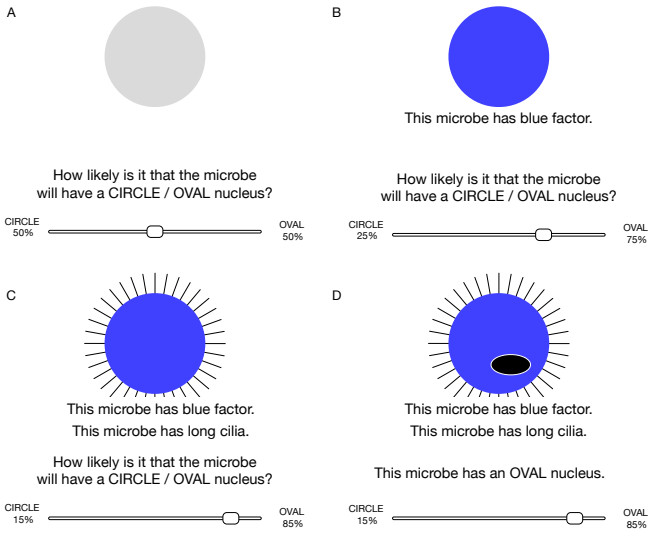

Figure 2: Trial-By-Trial Prediction and Feedback in Learning Phase.

After the learning phase, participants completed a test phase in which they made one-off judgments about microbe features. For example, participants might see a microbe with green cytoplasm and a round nucleus, and judge the likelihood that the microbe has short/long cilia. Participants made 11 judgments in this phase; in the current manuscript we focused on judgments of $P(X_2=1|Y=1, X_1=1)$ and $P(X_2=1|Y=1, X_1=0)$, and reported a few others as well.

To encourage participants to engage with the task, participants also earned bonuses if they guessed within 5% of the normative answer on five randomly-determined trials in the learning phase of each scenario. Each bonus was worth five cents; so it was possible to earn up to 50 cents in bonus pay. Participants were told the number of bonus trials, but were not told which trials were scored for bonuses.

Results

We first looked at other judgments that subjects made about these structures to confirm that they were indeed learning the probabilistic relations. Table 2 shows that the judgments were in the correct direction (above .5), confirming that they learned. (Remember that subjects were not told which states of the variables were correlated with each other, so they could not just have made these judgments in the right direction by the instructions alone or by guessing.) The fact that the judgments are not as strong as they should be normatively fits with many prior studies (Rottman & Hastie, 2014). The degree of weakness is similar to prior studies, suggesting that learning was not worse in the current study.

Table 2: Normative and observed probability judgments for basic inferences.

Dataset	Judgment	Norm	Mean (SD)
Mediation	$P(Y=1\|X_i=1)$	.83	.70 (.25)
Mediation	$P(X_i=1\|Y=1)$	.83	.73 (.22)
Mediation	$P(Y=1\|X_1=1, X_2=1)$	.95	.77 (.28)
CE1	$P(Y=1\|X_i=1)$	.77	.63 (.25)
CE1	$P(Y=1\|X_1=1, X_2=1)$	.89	.69 (.28)
CE2	$P(Y=1\|X_i=1)$	.75	.76 (.22)
CE2	$P(Y=1\|X_1=1, X_2=1)$	1.00	.72 (.32)

*Note: X_i denotes combined data for judgments that could apply to X_1 or X_2

We then focused more closely on the two key inferences. Means of the judgments are reported in Table 3, and histograms are presented in Figure 3. Overall, consistent with Rottman and Hastie (2016), many of the distributions are quite spread out, and often exhibit peaks at 0, .5, and 1. Participants were able to learn the basic relations; some of the average judgments were clearly in the correct direction.

Table 3: Normative and observed probability judgments for key inferences.

	$P(X_2=1\|Y=1, X_1=1)$		$P(X_2=1\|Y=1, X_1=0)$	
Dataset	Norm	Mean (SD)	Norm	Mean (SD)
Mediation	.83	.75 (.27)	.80	.52 (.31)
CE1	.47	.53 (.34)	.82	.62 (.29)
CE2	.17→.83	.66 →.33 (.32)	.63	.55 (.36)

Note: → means 'recoded as.'

For all inferential tests conducted below, when we compared two conditions, we conducted median split analyses because the distributions are not normal. We first calculated the pooled median judgment of the two conditions, and then tested whether the percent of judgments greater than the median was different across the two conditions, using a random intercept at the participant level and a random slope to allow the effect of the conditions to vary by participant.

Assessment of the Markov Assumption and Explaining Away

Though the main goal for this research was to compare $P(X_2=1|Y=1, X_1=1)$ and $P(X_2=1|Y=1, X_1=0)$[2] judgments in mediation vs. common effect structures, we first compared them within a structure, as they have typically been studied.

Focusing on the mediation structure, it is clear that the distribution for $P(X_2=1|Y=1, X_1=1)$ is higher than the $P(X_2=1|Y=1, X_1=0)$ distribution (Figure 3). We found a significant difference between the two judgments (B=2.41, SE=0.44, p<.001), evidence of a violation of the Markov assumption.[3] Importantly, this must be a learning effect, not a perceptual similarity effect or an effect from the instructions. Imagine that for one subject, X_2=1 means that the nucleus is an oval (not a circle), Y=1 means that the microbe is blue (not green), and X_1=1 means that the cilia are long (not short). The fact that they gave higher judgments when X_1=1 means that they learned the correlations between long cilia, blue, and oval, and that they overgeneralized these relations, believing that nucleus shape and cilia length were still correlated after controlling for the color. Another piece of supporting evidence is that the judgment of $P(X_2=1|Y=1, X_1=1)$, which was .75 on average, is relatively close to the normative value of .83 (at least on the right side of .50).

Did our subjects appropriately explain away? Within the Common Effect 1 structure, normatively the judgments of $P(X_2=1|Y=1, X_1=0)$ should have been higher by .35 than those for $P(X_2=1|Y=1, X_1=1)$. The difference was in the expected direction, but the mean difference was only .09; the median split analysis was not significant (B=-0.33, SE=0.29, p=.25). Within the Common Effect 2 structure, normatively the judgments of $P(X_2=1|Y=1, X_1=0)$ should have been higher by .46 than those for $P(X_2=1|Y=1, X_1=1)$, but the difference was only .03 (.55 vs. .52 in Table 3), and was not significant (B=0.51, SE =0.49, p=.30). This weak explaining away fits with prior research and with the cue consistency theory.

Comparing the Mediation and Common Effect Structures

Our primary goal was to compare $P(X_2=1|Y=1, X_1=1)$ and $P(X_2=1|Y=1, X_1=0)$ judgments across the two structures. We first compared the judgments of $P(X_2=1|Y=1, X_1=1)$. For the mediation structure this was normatively .83, and for the Common Effect 2 structure it was .17; we flipped[4] the judgments from the Common Effect 2 structure so they are comparable at .83. The judgments are considerably higher in the mediation structure (Figure 3). Due to order effects, we analyzed the data from the first scenario only, with the structure of the dataset as a between-subjects variable. Participants made higher (more accurate) judgments for mediation (M=.57) than Common Effect 2 datasets (M=.16), B=3.70, SE=1.07, p < .001. The mediation judgments were fairly good, but the common effect judgments were bimodal, with peaks at 0 and 1 (Figure 3).

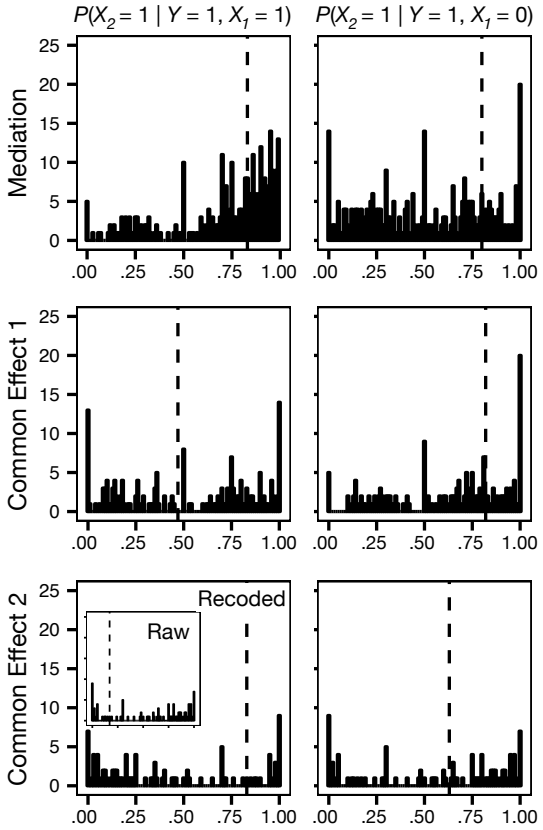

Figure 3: Histograms of participants' judgments of $P(X_2=1|Y=1, X_1=1)$ and $P(X_2=1|Y=1, X_1=0)$ for all three datasets. Dashed line indicates normative judgment.

To analyze the $P(X_2=1|Y=1, X_1=0)$ judgments, we compared the Mediation vs. the Common Effect 1 structure, for which the normative judgments are .80 and .82, respectively. The judgments were significantly higher (closer to normative) for the Common Effect 1 dataset (M=.60) than the mediation dataset (M=.45), B=-0.70, SE=0.24, p=.004.

[2] Because of the symmetrical nature of our datasets (e.g., $P(X_2=1|Y=1)=P(X_1=1|Y=1)$, we combined judgments about X_1 and X_2 and reported them as X_2. (E.g., $P(X_2|Y,X_1)$ includes $P(X_1|Y,X_2)$.)

[3] One limitation is that the normative judgments for $P(X_2=1|Y=1, X_1=1)$ is higher than $P(X_2=1|Y=1, X_1=0)$: .83 vs. .80. We doubt this slight difference caused the large differences in the judgments.

[4] If the variables were present vs. absent, then perhaps it would not make sense to flip them because the prospect theory probability weighting function is not symmetric around .5 (Tversky & Kahneman, 1992). However, in this study, the "lower" vs. "upper" half of the scales indicates higher/lower likelihoods for circle/oval nucleus, or long/ short cilia (Figure 2), so they are symmetric.

Across these two comparisons, sometimes judgments on the Common Effect structure look better than the mediation structure, and sometimes the reverse happens. There is no evidence that reasoning about one structure is uniquely bad.

Discussion

There were two main findings. First, we again found violations of the Markov assumption and insufficient explaining away. These findings fit with the cue consistency theory that subjects overestimate $P(X_2=1|Y=1, X_1=1)$ and underestimate $P(X_2=1|Y=1, X_1=0)$. However, in the current study the Markov violation arises *because* of learning, not in spite of it. Because the variables each have different states (e.g., long/short cilia, circle/oval nucleus), subjects must learn the correlations between the variables. They still violate the Markov assumption in the expected direction based on the bivariate relations, which implies that this effect is learned. It does not arise from prior knowledge, perceptual features, or instruction. Markov violations have also been found in studies that give subjects causal structures without learning data (e.g., Park & Sloman, 2013); causal knowledge and learning are each individually sufficient to yield this error.

Second, we found that judgments about common effect structures are sometimes better and sometimes worse than mediation structures; people are not uniquely bad at common effect structures, as one might think based on previous work.

It can be useful to consider which judgments are fairly normative, and which are especially bad, in order to problematize the bad judgments. The $P(X_2=1|Y=1, X_1=0)$ judgment was decent for the common effect, but was way too low for the mediation structure. This finding, that $P(X_2=1|Y=1, X_1=0)$ is more problematic than $P(X_2=1|Y=1, X_1=1)$ for violations of the Markov assumption, fits with Rottman and Hastie's (2016; Figure 3) results. In contrast, $P(X_2=1|Y=1, X_1=1)$ was especially problematic for the common effect. Previous research could not assess this; in prior studies (e.g., Rottman & Hastie, 2016), $P(X_2=1|Y=1, X_1=1)$ and $P(X_2=1|Y=1, X_1=0)$ were normatively in different parts of the probability scale and could not be compared.

These findings suggest that subjects' accuracy of inference is not determined by the inference question, nor the structure (the statistical relations), but by an interaction of the two.

One misinterpretation of these results, which we hope to head off, is that subjects did not learn in the CE2 condition. Though the distributions for CE2 in Figure 3 are spread out, there are clumps at 0 and 1, suggesting that some subjects believed the variables were related; if not, presumably they would answer near the middle of the scale to reflect uncertainty. Further, subjects learned the bivariate relations (Table 2). The difficulty is with the multivariate relations.

We have several future directions. First, a unique aspect of our design is that subjects made trial-by-trial predictions about each variable in the learning phase. We intend to develop learning models to explore how their difficulty with multivariate judgments could be tied to difficulty learning.

Second, it is unknown whether the current findings apply beyond certain probability ranges (around .83). The findings may change in different areas of the probability scale. Additionally, we only examined structures with symmetric parameters.

Third, Rehder and Waldmann (2017) presented evidence that multivariate inferences are *worse* when subjects know the causal structures that generated the data as opposed to not knowing the causal structures. It would be interesting to compare the current study to a similar one in which subjects know the causal structures; this would allow us to assess the degree of bias introduced by knowing the structure, and if different structures introduce different amounts of bias.

The current work explores how well people make inferences on three-variable causal structures, and allows for comparisons across structures by equating the normative judgments in the probability space. We found evidence that cue consistency effects can explain violations of the Markov assumption and insufficient explaining away. Further, we found that people are not worse at reasoning about some structures than others, but their accuracy also depends on the inference being made. This highlights the complexity of causal inference: people are sensitive to the learning data, but are also biased, and the bias interacts with the learning data.

Acknowledgements

This research was supported by NSF 1430439.

References

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

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology, 67*, 186-216.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*, 54-107.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264-314.

Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition, 45*, 245-260.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*, 109-139.

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology, 87*, 88-134.

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

Tversky, A., & Kahneman, D. (1992): Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297-323.