Causal Reasoning with Continuous Outcomes

Ahmad Azad Ab Rashid (abrashidaa@cardiff.ac.uk)
School of Psychology, Tower Building Park Place
Cardiff, CF10 3AT, UK

Marc J Buehner (buehnerm@cardiff.ac.uk)
School of Psychology, Tower Building Park Place
Cardiff, CF10 3AT, UK

Abstract

We describe an attempt to understand causal reasoning in situations where a binary cause produces a change on a continuous magnitude dimension. We consider established theories of binary probabilistic causal inference – ΔP and Power PC – and adapt them to continuous non-probabilistic outcomes. While ΔP describes causal strength as the difference of effect occurrence between the presence and absence of the cause, Power PC normalizes this difference with the effect base-rate to obtain a proportional measure of causal power, relative to the maximum possible strength. Two experiments compared the applicability of each approach by creating scenarios where binary probabilistic scenarios were directly mapped onto inference problems involving continuous magnitude dimensions. Results from counterfactual judgments tentatively indicate that people reason about causal relations with continuous outcomes by adopting a proportional approach when evaluation preventive causal powers, and a difference approach in generative scenarios.

Keywords: causal learning; continuous outcomes; reasoning; counterfactual.

Background

The capacity to learn about and represent causal knowledge is a fundamental aspect of cognition without which humans lose the ability to not only make predictions and decisions, but also to forecast, prepare and direct their behaviours towards achieving goals and fulfilling desires. Current research mostly focuses on causal relations involving binary events. Outside the lab, however, people do not only encounter binary events. In fact, we are more likely to be dealing with continuous variables: How much faster could I run if I lose 20 pounds of weight? How much weight would I gain if I ate cheeseburger everyday? How much sugar do I need to add to avoid over sweetening? These questions are daily examples of people's involvement with causal relations entailing continuous variables.

Binary causal relations involve a state change of a binary event (cause present/absent) to produce a change in another binary event (effect present/absent), but such simplicity is not the case for continuous variables. In a continuous causal scenario, a magnitude change of a continuous variable is produced by a magnitude change of another continuous variable. For example, in a binary relation, a state change of a cause could be flicking a switch from off to on which changes the status of a bulb from off to on. On the other hand, a continuous relation involves a change of a dial position to cause a change of luminosity from dimmer to brighter. Despite many daily-life examples of continuous variables, very few studies have been investigating causal judgment involving continuous variables (White, 2001). Here we are trying to find out how people acquire causal knowledge involving continuous variables?

Learning Framework: Difference or Proportion

Most theories of binary causal learning are rooted in Hume's empiricism (1739/1888): Causal knowledge is not explicitly available via sensory modalities but instead is inferred using the input received via them. One of Hume's cues to causation is contingency – i.e. the frequency of an effect and a cause co-occurring.

A longstanding model formalising contingency as an indicator of causal belief is ΔP, which calculates the difference of the probabilities of the effect in the presence vs. the absence of the cause (Jenkins & Ward, 1969):

$$\Delta P = P(e \mid c) - P(e \mid \neg c)$$

Consider these hypothetical scenarios involving the study of skin rash as a side effect of a new group of medicines on a group of forty patients. In scenario 1, none of them had a rash before taking medicine A, but 20 of them had rash after taking the medicine. In scenario 2, also none of them had rash before taking the medicine, but only 10 of them reported rash after taking medicine B. ΔP computes causal strength by considering the difference in relative frequencies of patients before and after taking the medicines, giving ΔP values of 0.50 and 0.25 respectively; hence concluding that medicine A has higher causal strength than medicine B to cause skin rash.

Consider another scenario 3 in which 20 of 40 patients already had skin rash even before taking medicine C, but the number of patients suffering with rash increased to 30 after taking the medicine. Applying ΔP in scenario 3 results in medicine C having a causal strength index of 0.25, which is similar to medicine B. However, studies have shown that despite having the same ΔP values, people tend to conclude that medicine C is more effective than medicine B in causing the rash (Cheng, 1997; Buehner, Cheng, & Clifford, 2003). This discrepancy is captured by another influential theory on causal learning: Power PC (Cheng, 1997).

Power PC argues that in addition to the difference causal strength is also influenced by the base-rate, P(e|¬c). Power

PC normalizes the difference with the base rate to obtain a proportional measure of causal power.

$$p_{gen} = \frac{\Delta P}{1 - P(E \mid \neg C)} \qquad p_{pre} = \frac{-\Delta P}{P(E \mid \neg C)}$$

Power PC has also been used to parameterise Bayesian models of causal learning (Griffiths & Tenenbaum, 2005) and is generally recognized as a rational account of causal strength.

Applying Power PC onto scenarios 2 and 3 results in having causal strength indexes of 0.25 and 0.50 for medicine B and C respectively. Unlike ΔP, this model therefore captures people's ability to provide normative responses. The key difference between ΔP and Power PC is that the former considers the *absolute difference* the cause makes to the occurrence of the effect, while the latter calculates the difference relative to the maximum causal change possible, and thus provides a *proportional* index of causal strength.

In the earlier scenarios, medicine B had the opportunity to cause skin rash in all 40 patients, and did so in 10 of them; in contrast, in the scenario involving medicine C, the medicine only had the opportunity to cause skin rash in 20 patients because the other 20 already had rash even before taking the medicine. From these 20 unaffected patients, medicine C managed to affect 10 of them to have skin rash. Therefore, Power PC suggests that for medicine B, the causal strength index is 0.25 because 10 out of 40 patients had rashes whereas for medicine C it is 0.50 because it caused rashes in 10 out of 20 (i.e. the initially unaffected) patients.

Moreover, the Power PC theory also tackles ceiling and floor effects. In another scenario where all 40 of the patients already had skin rash *before* taking medicine D, and all 40 still had skin rash *after* taking the medicine, ΔP for this scenario would be zero, suggesting that medicine D makes no difference to the occurrence of rash. A rational judgment, however, would be that the experiment is inconclusive with respect to generative causal power because medicine D had no opportunity to demonstrate its potential effectiveness, and thus the causal status of D is unknown. Wu and Cheng (1999) showed that reasoners indeed follow this logic, and withhold judgment in cases where causal power is unknowable. If Power PC is applied to this scenario, the equation is undefined (due to division by 0), which is consistent with both rational assessment and empirical results.

We highlighted the contrast between the difference and proportional perspectives of both theories because they will be relevant when considering approaches to continuous causation. Proportions can only be computed with respect to a reference limit. In binary probabilistic causation, the relevant limits are P(*e*) = 0 (the effect never happens) and P(*e*) = 1 (the effect always happens). These probabilities provide the upper limit of maximal causal effectiveness for preventive and generative causation, respectively, in a binary probabilistic framework: The maximum impact a preventor could have would be to reduce the probability of the effect to 0, while the maximum impact of a generator would be to raise it to 1. When considering causal changes to continuous outcome magnitudes, such natural limits are not necessarily present. While the maximum impact a preventor could have would still be to reduce the quantity of the effect to 0 magnitude, the maximum impact a generator could have might be unknown because it could keep on increasing the magnitude unless there is a known upper limit.

Study Scope

The central idea of this study was to investigate whether people reason about causal relations involving non-probabilistic continuous outcomes within a difference or proportional framework. Because of the wealth of prior works assessing the suitability of these approaches with respect to binary probabilistic causation, we wanted to create scenarios that afford a similar comparison between the two accounts. To this end, and as a first step on our quest, we only considered situations where a binary cause can produce a (deterministic) magnitude change on a continuous variable. This allowed us to set up situations that are one-to-one mappings of binary probabilistic causation to scenarios involving continuous outcomes. More specifically, in both cases the cause is still either present or absent, but instead of it resulting in a change of *probability* of the outcome, it now affects the *magnitude* of the outcome.

In probabilistic causation the (binary) cause results in a binary state-change across a group of entities; aggregating these state-changes across a sample results in an assessment of the change of probability of the effect brought about by the presence of the cause, which is of course a continuous variable bound between 0 and 1. In contrast, we considered changes of a continuous outcome magnitude in a single entity. This allowed us to preserve exactly the same structure as in probabilistic causal inference tasks. For example, a *probability* condition of P(e|c) = 0.75, which indicates that skin rash is present in 75 out of 100 patients given that all of them took the medicine, was mapped onto a *quantity* condition of Q(e|c) = 7.5 cm^2, indicating that 7.5 cm^2 of skin from an area of 10 cm^2 where the ointment was applied broke out with a rash.

In order to maximize comparability to binary probabilistic causation and preserve structural identity, our studies employed an artificial upper limit on a continuous scenario to serve as a reference for maximum causal effectiveness (see Method). Imposing such a limit allowed us to derive predictions not only for a difference based, but also for a proportional approach. Moreover, it afforded the opportunity for a more stringent test of the two approaches, by using different counterfactual scenarios to elicit causal judgments. More specifically, we asked one counterfactual question where the upper limit of causal effectiveness corresponded to the artificial limit in the learning phase, while another made reference to a higher limit, not previously experienced in the learning phase. If reasoners

approach causal inference problems involving continuous outcome magnitudes with a difference-based approach, changing the reference limit should have no impact on their predictions for causally induced magnitude change: All that would matter is the difference the cause made in the learning phase, regardless of the upper limit of causal effectiveness. In contrast, according to a proportional approach, reasoners would relate that difference to the maximum possible difference, and scale their predictions accordingly in the presence of a different limit.

Imagine that a government wants to test the efficacy of a 20 mph speed limit on traffic fatalities in residential areas. Community A serves as a pilot and fatalities are reduced from 20 per year before the trial to 10 per year after the trial. What would we predict if community B, which is larger, has more roads, and suffers from 50 fatalities a year, were to adopt the same program? According to a difference-based approach, we would predict that the program results in the same absolute reduction by 10, to result in 40 fatalities per year. The proportional approach would consider the maximum change possible in A (20) and would recognize that 10 corresponds to half of that. Consequently, it would predict a reduction from 50 to 25. We used a similar logic to compare difference to proportion based approaches.

Experiments

Participants

Thirty different undergraduates from Cardiff University's School of Psychology participated in each preventive and generative experiment in exchange for course credit.

Design and Procedure

Each participant worked on 15 conditions directly adapted from the binary probabilistic design of Experiment 1 in Buehner et. al. (2003). Each condition consisted of a pair of quantities of an effect in the presence vs. absence of the cause (see Table 1).

The generative experiment used a cover story that asked participants to imagine they were pharmaceutical consultants researching the side effects (skin rash) of synthetic substances in cosmetic creams. Fifteen different fictitious cosmetic creams corresponded to the 15 causal conditions in Table 1.

The cover story also described that the size of skin rash was measured before and after the application of the cream, and that some patients may develop skin rash even in the absence of any cosmetic products. Instructions stressed that each cream was applied to cover 10 cm^2 of a patient's back and that the base rate (rash before cream application) was also expressed with reference to this 10 cm^2 area. This served to impose an artificial limit of maximum causal efficacy – the cream could only create rash so as to cover the entire 10 cm^2 area.

A similar cover story was used for the preventive experiment, this time introducing ointments that relieve skin rash. Again, adopting the same 15 conditions, the story described a proper motivation on how allergic reaction would cause the skin rash to occur up to 10 cm^2 without any preventive measure, and on how the ointment would reduce the skin rash.

Table 1: Fifteen causal conditions for both experiments

Q(e|c)	Q(e|¬c)	|ΔQ|	Causal Power	
			Gen	Pre[1]
1.00	1.00	0.00	-	0.00
0.75	0.75	0.00	0.00	0.00
0.50	0.50	0.00	0.00	0.00
0.25	0.25	0.00	0.00	0.00
0.00	0.00	0.00	0.00	-
1.00	0.75	0.25	1.00	0.25
0.75	0.50	0.25	0.50	0.33
0.50	0.25	0.25	0.33	0.50
0.25	0.00	0.25	0.25	1.00
1.00	0.50	0.50	1.00	0.50
0.75	0.25	0.50	0.67	0.67
0.50	0.00	0.50	0.50	1.00
1.00	0.25	0.75	1.00	0.75
0.75	0.00	0.75	0.75	1.00
1.00	0.00	1.00	1.00	1.00

[1] Values of Q(e|c) and Q(e|¬c) are switched in preventive

After going through the cover story, participants were presented with 15 visual stimuli to correspond to the 15 conditions in a random order (see Figure 1). They then had to judge how strong the cause generates/prevents the effect by answering two counterfactual questions – one at a time. The two counterfactual questions were presented to correspond to two limits – a limit that was consistent with the cover story, and a higher limit.

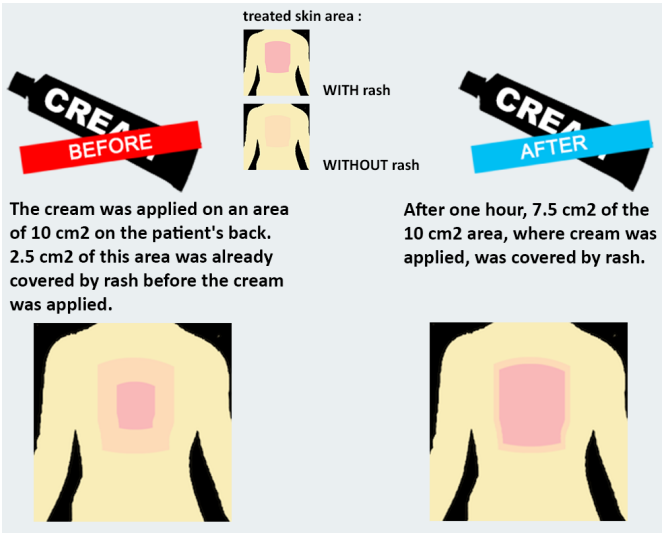

Figure 1: Sample Stimuli from the generative component

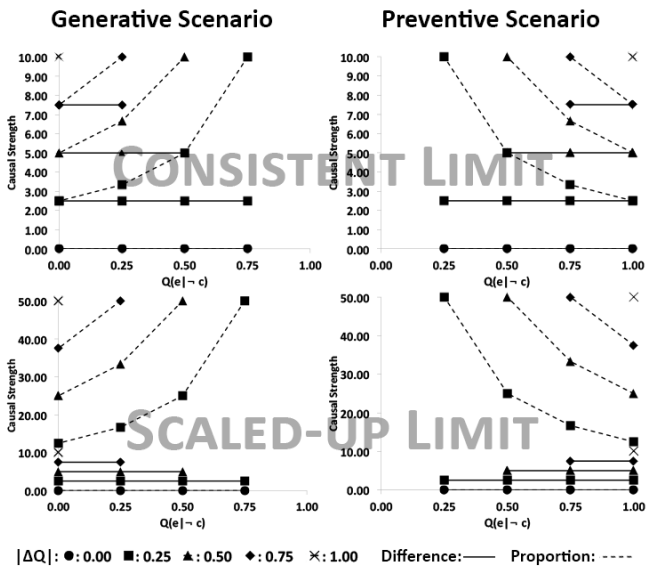

Figure 2: Power PC and ΔP Predictions of Causal Ratings.

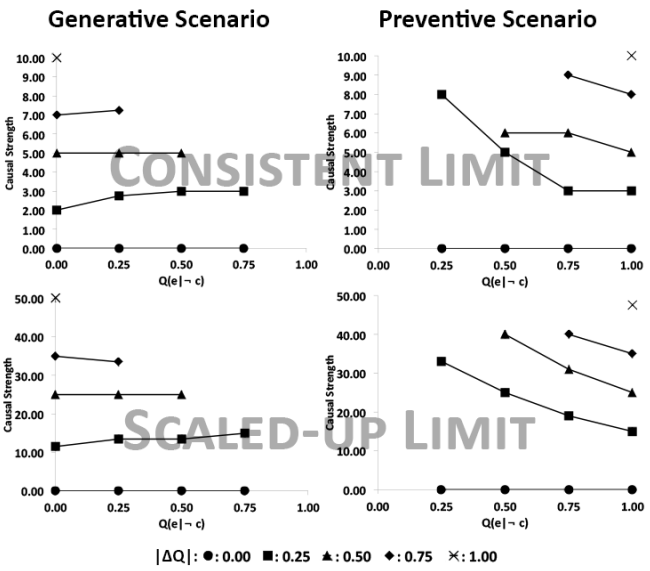

Figure 3: Medians of Counterfactual Responses.

The counterfactual question for the generative experiment was: *Now imagine a new patient who does not have any skin rash. If we applied this cream on the back of this patient to cover an area of 10 cm^2, how big would the area of skin rash be on this patient?* The exact same sentence was used for the second question except that the area (i.e. the limit) was changed to *50 cm^2*.

The counterfactual question for the preventive experiment was: *Now imagine a new allergy patient suffering from a rash of 10 cm^2. If we apply the ointment, how large would the area of rash be?* Similarly, the second question was exactly the same except for substituting the area with *50 cm^2*. Participants provided numerical responses using the keyboard.

Predictions

Figure 2 shows causal strength prediction plots for the 15 conditions, derived from difference based (ΔP) and proportional (Power PC) approaches (solid and dashed lines respectively). Causal conditions that have identical ΔQ values are linked together and plotted against the base-rate.

To allow comparisons both with previous literature, and across the two limit scenarios, these predictions were plotted with respect to the value of the limits tested. Since the maximum area of skin rash is 10 cm^2 in the consistent-limit scenario, the maximum power in the prediction has been set to 10 as well. In contrast, in the scaled-up limit scenario, the maximum power in the prediction has been set to be at 50 to match up with the maximum rash area of 50 cm^2.

Participants' judgments were analogously converted: For instance, an area judgment of 10 cm^2 in the consistent-limit scenario was converted into a causal rating of 10 in the generative, and a causal rating of 0 in the preventive experiment.

More specifically, we subtracted the counterfactual response given by the participant from the relevant upper limit. This conversion was made on the judgments to reflect that an increase of affected skin area would indicate an increase of causal power when considering generative causes, while larger predicted skin areas would indicate weaker causal powers when considering preventive relations.

The absolute difference approach predicts that causal strength is unaffected by increments of base-rate, and that causal ratings vary only as a function of ΔP. Furthermore, a strict interpretation of difference approach would suggest that the same difference is then applied to a different context, involving a higher upper limit. Consequently, prediction plots for the difference approach remain within the range of 0 to 10, across both the consistent-limit and scaled-up limit scenarios.

The proportional approach, on the other hand, predicts a consistent influence of base-rate onto causal ratings in both limit cases, which varies depending on whether generative or preventive powers are assessed. Despite having the same non-zero difference values (i.e. ΔP), in the generative scenario causal ratings should increase as the base rate increases. The reverse pattern is predicted in preventive scenarios. These influences of base rate, however, are not predicted for when the difference value is zero, and causal ratings should remain at zero for both generative and preventive cases. In addition, the proportional approach also dictates that counterfactual causal ratings are scaled up in line with a higher limit.

Results

Kolmogorov-Smirnov test showed that judgments were non-normally distributed. Consequently, Figure 3 plots median

judgments, and statistical analysis was based on non-parametric tests.

A qualitative inspection of the generative results in Figure 3 suggests that judgments correspond more to difference than proportional approach predictions. In the consistent-limit scenario, apart from the conditions involving ΔQ = 0.25, the judgments for other ΔQ values are relatively flat at the predicted difference values, suggesting a minimal influence of base-rate.

This minimal influence of base-rate is also evident on causal judgments in the scaled-up limit scenario. In this scenario, judgements from conditions involving identical values of ΔQ are also relatively consistent at the difference values, even though a small indication of a positive trend is observed in the ΔQ = 0.25 case. Even though the minimal influence of base-rate influence is in line with a difference account, generative judgments violate its other significant property: They vary from 0 up to 50, instead of 10. We will discuss this in the next section.

Qualitatively inspecting the preventive results in Figure 3 suggests they fit well with proportional approach. In both limit scenarios, the contingent cases indicate the influence of the base-rates. Instead of remaining constant at the difference values, the judgments decrease as the base-rate increases. Moreover, for the non-contingent cases, judgments also follow proportional predictions, in that they stay at zero despite a change of the base-rate. Even though there is an indication of a non-normative trend in the consistent-limit scenario when ΔQ = 0.25, in general, the preventive judgments seem to have followed proportional predictions, both with a consistent and inconsistent limit.

Statistical Analysis (Generative) Nonparametric Friedman's ANOVA was used to determine the main effect of the base-rate for every ΔQ value.

Analysis of ratings from the consistent-limit case found a significant effect of base-rate when ΔQ = 0, X_F^2 (14) = 14.750, $p < .05$ and ΔQ = 0.25, X_F^2 (14) = 10.545, $p < .05$. The analysis does not show any significant effect of base-rate when ΔQ = 0.50, X_F^2 (14) = 0.347, $p > .05$ and ΔQ = 0.75, X_F^2 (14) = 1.190, $p > .05$.

Unlike in the consistent-limit case, analysis of the scaled-up limit scenario shows a significant effect of base-rate only when ΔQ = 0.25, X_F^2 (14) = 7.978, $p < .05$. No significant effects of base-rate are found when ΔQ = 0, X_F^2 (14) = 6.681, $p > .005$; ΔQ = 0.50, X_F^2 (14) = 1.357, $p > .005$; and ΔQ = 0.75, X_F^2 (14) = 1.087, $p > .005$.

Surprisingly, the statistical test indicates an effect of the base rate in the non-contingent case of consistent-limit scenario, despite an observation of a flat line in Figure 3. Inspection of the data distribution in these conditions (Figure 4) reveals three noteworthy points: i) the modal response is 0 in all cases, ii) a minority of participants give a non-normative non-zero response, iii) this minority of participants appears to exhibit an outcome density bias (Buehner, Cheng, & Clifford, 2003). Because the Friedman Test ignores ties, the significant result in ΔQ = 0 condition is thus driven by this minority of participants.

Statistical Analysis (Preventive) In the consistent-limit scenario, no significant effect of base-rate was found when ΔQ = 0 , X_F^2 (14) = 4.500, $p > .05$. However, significant effects of base-rate were obtained when ΔQ = 0.25, X_F^2 (14) = 57.854, $p < .05$; ΔQ = 0.50, X_F^2 (14) = 15.892, $p < .05$; and ΔQ = 0.75 as well, X_F^2 (14) = 9.783, $p < .05$.

Similar trends were observed in the scaled-up limit scenario. The analysis shows no significant effect of base-rate when ΔQ = 0, X_F^2 (14) = 1.222, $p > .05$. Again, significant base-rate effects are found when ΔQ = 0.25, X_F^2 (14) = 27.931, $p < .05$; ΔQ = 0.50, X_F^2 (14) = 12.302, $p < .05$; ΔQ = 0.75, X_F^2 (14) = 3.846, $p < .05$.

As with the generative scenario, non-contingent conditions uniformly elicited a median and modal response of zero. While there was also a minority of participants who deviated from this normative assessment, judgments from these participants did not display any systematic patterns. More specifically, unlike in the generative scenario, there was no evidence of an outcome density bias, even in the minority of non-normative judgments.

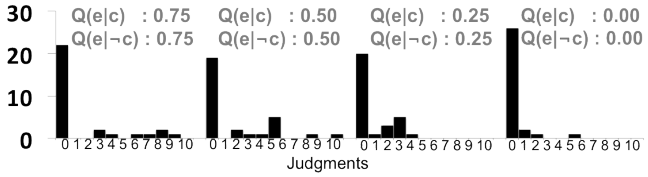

Figure 4: Judgment Distributions of non-contingent Conditions (ΔQ = 0) in the consistent-limit scenario

Discussion

Overall, our results seem to suggest that when people reason about continuous outcomes, they do so within a proportional framework, if the context is one of preventive causation, i.e. the goal is to reduce the outcome magnitude. However, if the context involves increasing the outcome magnitude (generative causation), people seem to focus on the difference the cause makes, without normalizing this difference to an upper limit, even when the task clearly implies such a limit. Interestingly, people then do not adhere to the absolute difference a cause makes in a given context, but instead scale up this difference, where appropriate, in different scenarios.

For instance, in the condition when Q(e|c) = 1.00 and Q(e|¬c) = 0.25, participants learned that a skin area of 2.5 cm^2 was covered with of rash before the application of the cream, and that applying the cream to an area of 10 cm^2 resulted in that entire area breaking out with rash. They considered the difference the cream made, and concluded that its application increases the area of rash by 7.5 cm^2 when applied to 10cm^2 of skin of a patient who does not yet suffer from rash. Had they taken the proportional approach, they would have concluded that this cream is maximally

effective in producing rash, and applying it to an area of 10cm^2 of healthy skin would lead it all of it to break out with rash. When they were asked to transfer their knowledge to a different scenario, where the cream was applied to 50cm^2 of healthy skin, they took the difference (7.5 cm^2) and scaled it up to this new area, concluding that 33 cm^2 (i.e. nearly 37.5cm^2) of the 50 cm^2 will break out with rash.

Inspection of Figure 3 shows that participants were relatively consistent in scaling up their counterfactual judgments across all the generative conditions: a factor of approximately 5 emerges. This suggests that participants indeed scaled up their judgments from one context to the other, rather than merely considering the difference, as suggested by a strict interpretation of a difference-based approach. It appears then that people were aware of the upper limit we imposed on our scenarios, and scaled their judgments up accordingly in both preventive and generative situations. However, the judgments they formed were based on proportions only for preventive contexts, and on differences in generative contexts.

One tempting conclusion might be that perhaps our generative cover story might simply have failed to instill a clear sense of an upper limit in the learning phase, despite our best efforts to do so. After all, even when cream is applied to only to 10 cm^2, it is still feasible for a rash to occur in a larger area than that. In contrast, the preventive scenarios were not hampered this way – the natural upper limit of preventive causation is always 0: No treatment could reduce rash to less than an area of 0 cm^2. However, we have conducted studies with other generative contexts, involving continuous outcome magnitudes that definitely do have clear and unambiguous upper limits (such as relative humidity in the atmosphere), and the results mirror those reported here: People largely adopt a difference-based approach when evaluating generative causal influence.

Conclusions

The work reported here represents the beginning of a quest to chart the waters of continuous causal inference. We have taken a cautious approach and created situations that are structurally identical to conventional binary probabilistic causal inference. We knew that doing so would limit the ecological validity of our results. After all, most causes are continuous variables themselves, influencing continuous outcome magnitudes. However, our goal here was a proof of concept: We wanted to measure people's inferences about causal change to continuous outcomes under ideal conditions and with clear explicit upper limits (which are not always present in the world). If under these conditions, inferences followed patterns similar to those observed in probabilistic causal inference, this might suggests that a fruitful avenue to pursue might be to try and adapt theories and models from binary probabilistic causal inference to inference about continuous causation.

Tentatively, we would conclude that people's inference patterns do correspond to what we know about probabilistic causal inference. Deviations from normative models are found frequently also in probabilistic causal inference (e.g. Lober & Shanks, 2000), although sometimes such deviations seem to reflect ambiguities in the task demands. And indeed perhaps the non-normative results of our generative experiment may be due to such ambiguities. We are currently addressing this with follow-up studies. For example, we have not considered the reliability of the information on which participants base their judgments. Bayesian models of causal inference (e.g. Griffiths & Tenenbaum, 2005) consider both the strength of a causal relation (as indexed by power PC), as well as the reliability of the information (as indexed by the sample size, or the effective sample size). For simplicity, and to ensure the one-to-one mapping to probabilistic causation, our study involved only single entities (i.e. one patient per treatment).

In future work, we hope to consider not only multiple instances of continuous outcome change from the same cause, but also to begin working with causes that are in themselves continuous variables.

Acknowledgments

This work has been carried out as part of a PhD by the first author, supervised by the second author. We acknowledge funding from Malaysian Institute of Road Safety Research (MIROS) to support this study.

References

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 29(6), 1119.

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

Griffiths A, Tenenbaum J.B. (2005). Structure and strength in causal induction. *Cognitive Psychology*. 51, 334–84.

Hume, D. (1888). *Hume's treatise of human nature* (L. A. Selby-Bigge, Ed.). Oxford, UK: Clarendon Press. (Original work published 1739).

Jenkins, H., & Ward, W. (1965). Judgment of contingencies between responses and outcomes. *Psychological Monographs, 7,* 1-17.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review, 107*(1), 195-212.

White, P. A. (2001). Causal judgment about relations between multilevel variables. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* 27, 499-513.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of gener- ative and preventive causal powers. *Psychological Science*, 10, 92–9.