Causal Reasoning with Continuous Outcomes

Ahmad Azad Ab. Rashid

Cardiff University

This dissertation is submitted for the degree of PhD.

2015

Abstract

Twenty experiments investigated how people reason about causal relations where a binary cause (present/absent) influences the continuous magnitude of a target outcome. The experimental design was based on a conceptual mapping of probabilistic influences in binary causation to deterministic influences on continuous effects. Doing so preserved the computational properties related to binary causation, and allowed me to test applicability of well-established causal reasoning strategies in continuous causation. The investigation employed three methods: the first one involved asking participants the standard causal questions on strength rating; the second method asked other participants to make judgments in accordance to counterfactual questions; and the third method required participants to identify the direction candidate cause influenced effect magnitude. Results reveal that when reasoning about binary causes that reduce a continuous outcome magnitude, the support is for proportional reasoning approach, which is conceptually equivalent to the Power PC theory of binary causation. When reasoning about causes that increase a continuous magnitude, however, the results did not converge to any prominent strategy because of various moderating factors. Moreover, under certain circumstances, reasonsers also appear to adopt a strategy based on a multiplicative reasoning, which has not been documented in the literature before. The evidently low consistency of results within participant and within condition across experiments suggests that neither approach properly explains this type of reasoning.

Preface

In the name of Allah, the Most Gracious, the Most Merciful. All praise is due to Allah, who alone has the causal power. With the incapacity of me to understand the whole picture of causality, the works presented here are only within the sublunary substance and framework.

I would like to express appreciation to my sponsor for giving me an opportunity to complete this study, and to my supervisor Dr. Marc Buehner for his advice, guidance, patience, understanding, and occasional poking. Also, my gratitude goes to my family: my father and mother, who always pray for my best; my siblings for constant encouragement; my wife (who joined me in the second year of study) for her love, support and comfort; and to my first daughter (who came to this world in the third year of study) for her everything. Next, I would like to thank my friends for their assistance and advice. Lastly, but not the least, I would also like to thank the staff and postgraduate students at Cardiff University for their helpful comments and criticisms.

Terima kasih!

Azad

Table of Contents

Chapter 0: Overview of Thesis

This thesis begins with an introduction to causation, and the importance of causal knowledge before raising the main question: how do people acquire causal knowledge? Because of the scarcity of previous work on continuous causation, I explored the binary causation literature to identify the two most influential frameworks as a starting point for this journey: ΔP and Power PC theory. Following this, the next question was to investigate whether these probabilistic binary reasoning strategies are applicable in explaining causal relations entailing continuous outcomes. If they do, which strategy would better capture this type of reasoning?

The first empirical chapter (Chapter 2) opens the investigation with two experiments: The first one compared binary and continuous causation in both generative and preventive scenario, while the second experiment aimed to study the influence of limit saliency. In both experiments, I begin with the simplest way, by asking participants to report the extent of the candidate cause in producing/inhibiting the effect magnitude. I continued the investigation with a hypothetical judgment approach in Chapter 3.

Chapter 3 is the second empirical chapter, containing ten experiments adopting a "counterfactual" approach as Buehner, Cheng and Clifford (2003) suggested. Because they argued that this approach would better tap into causal power, i.e. following Power PC theory, it would be insightful to see whether the proportion strategy, which was adapted from Power PC, could dominate reasoning with continuous outcomes.

I presented another eight experiments in Chapter 4 adapting the implicit judgment approach from Liljeholm and Cheng (2007). Unlike the previous two chapters, the approach in this chapter did not ask participants to explicitly measure causal *strength*, but instead only

required them to identify the *direction* in which a candidate cause influences the effect magnitude.

The final chapter summarises the results, and highlights factors that were moderating participants' judgments in all experiments. In line with Perales and Shanks (2008) argument, variations of the results could be attributed to participants' judgments in accordance to demands of the task at hand. To test this, I revisited the results in the light of Evidence Integration rule as Perales and Shanks proposed. Further, I concluded the thesis and also included my thoughts on potential future research on causal learning between continuous variables.

Following the main content of the thesis are appendices: Appendix A lists all cover stories of all experiments, Appendix B describes the determination of bin size for the histogram analysis, and the gamut for the tendency analysis of Chapter 3, and Appendix C discusses the relation between the proportion and multiplication strategies.

Chapter 1: Background and Study Approach

1.0 Causality and Causal Learning

Causal knowledge is central to cognition (Sloman, 2005) and a prerequisite for effective reasoning and problem solving (Newsome, 2003), and without which we are detached from our surroundings; it is central to make predictions, decisions and judgments, to interact and navigate within the world in order to fulfil motivations and goals, or to avoid hazards and harmful situations. The description of what a non-causal world would be, which Cheng and Buehner (2012) describe in the beginning of their article, puts further emphasis on the role of causal knowledge in our daily activities, especially when we are making predictions. Cheng and Buehner further highlight that "[c]ausation, and only causation, licenses the prediction of the consequences of actions".

How does a person come to have knowledge that something causes another thing? Traditionally, behaviourists viewed learning more as *reflex-oriented* acquisition of links between stimuli and response or behaviour and outcomes. Later, another view postulated that the main purpose of learning is to discover the *causal texture* of the world (Tolman and Brunswik 1935). This view traces back to David Hume (1739/1888) who argued that causal knowledge is not readily accessible using the sensory modalities; instead, people use the input acquired via them (e.g. observation of events occurring) to infer causal relations.

1.0.1 Causal Learning about Binary Relations

Current research mostly focuses on causal relations involving binary events (for an overview see Cheng & Buehner, 2012; Perales & Catena, 2006; Perales & Shanks, 2007). 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). 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.

Most theories of binary causal learning are rooted in Hume's empiricism, and stemmed from contingency (i.e. the frequency of an effect and a candidate cause co-occurring). The more often a candidate cause and an effect co-occur, the more likely for people to induce the cause to produce the effect. The simplest representation to capture relation and contingency information for binary causation is via a 2 x 2 contingency table (see Table 1.1)

Table 1.1: Standard contingency table that captures binary causal relations

	Effect present (e)	Effect absent (¬e)
Cause present (c)	a	b
Cause absent (¬c)	c	d

Note: a, b, c and d are frequencies respectively correspond to events where both cause and effect are present, only cause is present, only effect is present, and both cause and effect are absent.

In the table, a, b, c, and d, respectively corresponds to frequency of events when both the effect and candidate cause are present, when only candidate cause is present in the absence of effect, when only effect is present in the absence of candidate cause, and when both effect and candidate cause are absent.

ΔP Rule. A longstanding model formalising contingency as an index of causal belief is ΔP (Jenkins & Ward, 1965; Ward & Jenkins, 1969). This model argues that people estimate causal belief by considering the difference of the conditional probabilities of the effect in the

presence (P(e|c)) versus in the absence (P(e|¬c)) of the cause: ΔP = P(e|c) – P(e|¬c). These probabilities can be estimated from the frequencies information of the 2 x 2 contingency table: $P(e|c) = \frac{a}{a+b}$ and $P(e|\neg c) = \frac{c}{c+d}$. The model is a statistical representation of one-way contingency of an event on another (Allan, 1980).

Consider these hypothetical scenarios involving the study of influence of minerals on algae growth on a group of 100 pools. In scenario 1, 30 of the pools already had algae growth before receiving treatment with mineral A, and 65 of them had algae growth after receiving the treatment. In scenario 2, none of the pools had algae growth before receiving treatment with mineral B, and 50 of them had algae growth after receiving the treatment. ΔP computes causal strength by considering the difference in relative frequencies of pools before and after treatment with minerals, giving ΔP values of $\frac{65}{100} - \frac{30}{100} = \frac{35}{100} = 0.35$ and $\frac{50}{100} - \frac{0}{100} = \frac{50}{100} = 0.50$ respectively, hence suggesting that mineral B has higher causal strength than mineral A to cause algae growth.

Power PC Theory. Consider a third scenario in which 25 of 100 pools already had algae growth before receiving treatment with mineral C, and the number of pools covered with algae increased to 75 after receiving the treatment. Applying ΔP in scenario 3 results in mineral C having a causal strength index of 0.50 ($\frac{75}{100} - \frac{25}{100} = \frac{50}{100}$), which is the same as mineral B. Studies involving scenarios such as these, however, have shown that despite having the same ΔP values, people tend to conclude that mineral C is more effective than mineral B in causing algae growth (Cheng, 1997; Buehner et al., 2003). This reasoning discrepancy is captured by another influential theory of causal learning: the Power PC Theory (Cheng, 1997).

The Power PC Theory claims that ΔP on its own is not a useful index of causal strength, because it tracks the proximal stimulus (the observable contingency), when the goal of causal induction is to track the distal stimulus (the unobservable causal power). The concept of

unobservable causal power traces its root to Kant (1781/1965), who proposed that there exists a priori knowledge or framework that people refer to in order to interpret information during causal induction.

Cheng explains causal power by purporting laws and theories in science: In science, "laws, … which deal with observable properties, are often explained by theories, which posit unobservable entities". For instance, when *Boyle's Law* demonstrates that the absolute pressure of an ideal gas is inversely proportional to volume it occupies and remains unchanged within a closed system, *kinetic theory of gases* explains this phenomenon by positing gas as tiny particles that are constantly moving at random. Because their speed is proportional to temperature, when these particles bombard the container walls, the gas law yields. Cheng further argues that "causal power is to covariation as the kinetic theory of gases is to Boyle's law."

Computationally, Power PC Theory tracks causal power by normalising ΔP with the base rate (i.e. probability of effect presence in the absence of the candidate cause). The resultant proportional measure of causal power for generative and preventive (strength index) candidate cause respectively is $\frac{\Delta \mathrm{P}}{1-P(e|\sim c)}$ and $\frac{-\Delta \mathrm{P}}{P(e|\sim c)}$.

Applying Power PC to scenarios 2 and 3 results in having causal strength indices of 0.50 and 0.66 for mineral B and C respectively: In the earlier scenarios, mineral B had the opportunity to cause algae growth in all 100 pools, and did so in 50 of them; in contrast, in the scenario involving mineral C, the mineral only had the opportunity to cause algae growth in 75 pools because the other 25 already had growth even before treated with the mineral. From these 75 unaffected pools, mineral C managed to affect 50 of them to have growth. Therefore, Power PC suggests that for mineral B, the causal power is 0.50 because 50 out of 100 pools had growth

whereas for mineral C it is 0.66 because it caused growth in 50 out of 75 (i.e. the initially unaffected) pools.

Moreover, the Power PC theory also addresses ceiling and floor effects. Imagine another scenario 4 where all 100 of the pools already had algae growth before receiving treatment with mineral D, and all 100 still had algae growth after receiving the treatment, ΔP for this scenario would be zero, suggesting that mineral D makes no difference to the algae growth. A rational judgment, however, would be that the experiment is inconclusive with respect to generative causal power because mineral D had no opportunity to demonstrate its potential effectiveness, and thus the causal status of mineral D is unknown. Wu and Cheng (1999) demonstrated that reasoners 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. Note that when considering preventive power for this scenario, Power PC agrees with ΔP in suggesting that the treatment is ineffective.

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. I highlighted the contrast between the difference and proportional perspectives of both theories because they will be relevant when considering potential approaches to continuous causation in the following section.

1.0.2 Continuous Causation

Over the past 30 or so years, researchers focused on identifying mechanisms of acquiring causal knowledge from covariation information. As highlighted in previous sections, frameworks for binary causal reasoning were mainly focusing on contingency between the candidate cause

and effect, specifically, on information in the 2 x 2 contingency table (see Table 1.1) as the focal of research interest.

Via their sensory modalities, people do not only observe and acquire information about the presence and absence of events, they also encounter causal relations involving continuous variables: How much faster could I run if I lose 20 pounds of weight? How much weight would I gain if I ate a cheeseburger everyday? How much sugar do I need to add to avoid over sweetening? How much algae would grow if I pour a gallon of phosphorus into a lake? These questions are examples of people's involvement with causal relations entailing continuous variables.

In contrast to binary causal relations, continuous causal relations involve a magnitude change of a continuous variable to produce a magnitude change in another continuous variable. An example of a continuous relation can be a change of a dial position to cause a change of luminosity from dimmer to brighter. Despite many related daily-life activities, very few studies have been investigating causal judgment involving continuous variables.

Causal learning about continuous relations. A simple continuous causal relation involves changes between a continuous cause and a continuous effect. The cause increases the magnitude of the effect in generative scenarios, but reduces it in preventive scenarios. To my knowledge, no studies have examined causal learning of relations between continuous variables except Young and Cole (2012).

Young and Cole conducted the study using a video game paradigm where participants had to determine a true cause among three other alternative causes. These causes, represented as enemy creatures, moved around in elliptical motion surrounding of an enemy detector that emitted tone at certain pitch (continuous effect) as a function of the creature's proximity from the

detector (continuous cause). In addition to varying the relations between these continuous variables, Young and Cole also manipulated the probability of the detection (i.e. probabilistic effect), as well as the range of sound pitches. In the first experiment, they found that participants' accuracy performance was independent of probability of effect, and pitch range. The participants did score higher than chance but at marginally (47% versus 33% chance). In the second experiment, they simplified the task by fixing the probability of effect to one (always happen), and by introducing a fixed detection proximity radius (no tone sounded outside this radius). The results of this experiment were better: Participants' accuracy was dependent on the pitch range, and higher than in the previous experiment, at 70%.

Further, Young and Cole claimed that participants' better performance in the latter experiment was because of the presence of fixed detection proximity radius, and not because of the fixed probability. This claim was based on their unpublished findings. While this work of Young and Cole was the first published evidence of the ability of people to recognize and responsed to the causal relation between two continuous variables, it does not consider of any theoretical framework on how people learn causal relation entailing continuous variables.

Given the scarcity of previous works in this area I begun this journey by a smaller step: Specifically, unlike Young and Cole's work, I focus on the deterministic relation between a binary cause with a continuous effect. In other words, all causal relations of this study were focusing on changing of effect continuous magnitudes corresponding to binary state change of causes. However, unlike Young and Cole, my work included theoretical frameworks adapted from binary causation with the aim to explore whether they are applicable for continuous outcomes. The next section contains a more detailed explanation of this approach.

1.1 Study Approach

The objective of this study was to investigate the reasoning strategies people adopt when contemplating continuous causation. Specifically, I wanted to determine whether reasoning approaches relevant to learning probabilistic binary causal relations are also applicable to continuous causation. I concentrated my efforts on the difference concept of ΔP and the proportion concept of Power PC. To do this, and as a first step on my quest, I only considered situations where a binary cause produces a (deterministic) magnitude change on a continuous variable. This allowed me 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 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, I considered changes of a continuous outcome magnitude in a single entity so that I could preserve exactly the same cognitive structure as in probabilistic causal inference tasks. As an example, a probability condition of $P(e|c) = 0.75$, which indicates that algae growth is present in 75 out of 100 pools given that all of them were treated with the mineral, was mapped onto a quantity condition of $Q(e|c) = 75\ m^2$, indicating that 75 m^2 of a pool surface area was covered with algae. In short, I directly mapped a change of probability to a change of quantity of a continuous variable.

I employed this direct mapping approach to study the relevance of the core concepts behind the ΔP rule (difference) and Power PC (proportion) in relation to continuous outcomes.

Let us review these concepts in the context of continuous outcomes using scenario 3 as an example. A direct mapping of this scenario onto a continuous outcome (I refer to this mapped scenario 3 as scenario 3C) results in mineral C producing algae growth in terms of surface area of a single pool. Before treatment with mineral C algae covered 25 m^2 of the surface area of the pool, and this increased to 75 m^2 after treatment with the mineral. Adopting the difference concept of the ΔP rule, the efficacy of mineral C can be inferred as 50 m^2 algae growth. In contrast, by adopting the proportion concept of Power PC, the efficacy of mineral C can be computed by taking a proportion of the difference, 50 m^2, relative to the maximum effect the mineral could have produced. The maximum efficacy, however, can only be computed if a limit of outcome magnitude is clearly defined. To address this problem, I imposed artificial limits in my experiments. For this scenario, because the continuous variable that I am considering is the area of the pool surface, the size of the pool defined the maximum area of the algae growth. If I defined the size (and hence the limit) as 100 m^2, then the efficacy of mineral C using a proportion approach is 0.666, which is the ratio of 50 m^2 to 75 m^2 (100 m^2 - 25 m^2). This 0.666 efficacy of mineral C is the same as its binary counterpart in scenario 3.

Defining an upper limit of causal efficacy is only an issue for generative scenarios. As per the above example, mineral C can keep producing the magnitude of the effect to, theoretically, an infinite amount of square meter of algae if there is an infinite area of water surface. In the absence of knowing the theoretically possible maximum efficacy of mineral C (in this case the size of the surface area), the proportion strategy is incomputable. On the other hand, there is always a limit in preventive scenarios because a magnitude of zero exists naturally in any situation (as long as the variable of interest is measurable on a ratio scale). If I consider mineral

C as having a preventive capacity, then the maximum efficacy of preventive power can always be implied as a reduction to 0 m^2, i.e. the complete wipe-out of algae growth in the pool.[1]

1.2 Reasoning Strategies

To demonstrate the relevance of difference and proportion strategies in reasoning about causal relations involving continuous outcomes, let us go through the following scenario 5. Imagine a new pool with a surface area of 100 m^2, and algae have already covered 10 m^2 of the surface. If I ask a reasoner: How big would the area covered by algae be, if the same mineral C from scenario 3C were administered into the pool? I could infer his or her reasoning strategy based on the answer given as follows: A reasoner who strictly uses a difference strategy should predict algae growth to an area of 60 m^2. This is because, in scenario 3C, he or she has learned that mineral C has a difference-based efficacy of 50 m^2. Therefore, given the initial area of growth was already 10 m^2, the final area after the treatment should then be 60 m^2. On the other hand, a reasoner who adopts a proportion strategy would firstly compute the potential maximum efficacy mineral C could have in this scenario as 90 m^2 (100 m^2 – 10 m^2). Because he or she learned in scenario 3C that mineral C has an efficacy of 66.6%, in this scenario, it should also produce algae according to this proportion. Therefore, given that there are 90 m^2 surface area remaining on which the mineral could cause growth, the reasoner would predict that it would do so on 66.6% of this area (i.e. 60 m^2). Adding this to the already covered 10 m^2 yields a predicted total surface area of 70 m^2 covered with algae after treatment.

Besides 60 m^2 and 70 m^2, another plausible prediction for scenario 5 is 30 m^2: Reasoners might consider the efficacy of mineral C in terms of its capacity to multiply the covered surface

[1] In binary probabilistic causation, as a comparison, the relevant limits are the probability of the effect, P(e) = 1 (the effect always happens) in generative scenarios, and P(e) = 0 (the effect never happens) in preventive scenarios. These probabilities provide the upper limit of maximal causal effectiveness: The maximum impact a

area relative to before treatment. A reasoner who adopts this strategy learnt that in scenario 3C mineral C tripled the area of algae growth from 25 m^2 to 75 m^2. Therefore in scenario 5, with the same efficacy, mineral C should also triple the algae growth from 10 m^2 to 30 m^2. I refer to this approach as the multiplication strategy. I am unaware of any discussions of a multiplication strategy in the literature, but my pilot studies suggested that some reasoners might adopt this approach. Even though the multiplication strategy surfaced only in the generative scenarios of the pilot studies, it makes sense in preventive scenario as well; hence the inclusion of this novel strategy alongside difference and proportion strategies in both generative and preventive scenarios of this investigation.

The proportion strategy, unlike the other two strategies, requires reasoners to hold certain assumptions for it to be applicable (for an elaborate discussion on assumptions of Power PC, from which the proportion strategy was adapted, see Cheng, 1997). One of the critical assumptions is the assumption of independence: Reasoners assume that the candidate cause influences the effect magnitude irrespective of the influence of background causes. In other words, the influence of background causes is assumed to be constant in the 'before' and 'after' (or 'with' and 'without') treatment cases. For example, for proportion to work, reasoners have to assume that the already existing influence of background causes on growth does not affect the influence that mineral C has on growth.

Both the difference and multiplication strategies violate assumption of independence. People who adopt a difference strategy compute the influence of a candidate cause onto the effect as if the background causes have completed (and stopped) their influence on the effect magnitude before allowing the candidate cause to produce the change only onto the remaining

---

preventor could have would be to reduce the probability of the effect to 0, while the maximum impact of a generator

unaffected magnitudes. In other words, difference based reasoners think that the influence of the candidate cause and background causes onto the effect are mutually exclusive, which opposes the assumption of independence. For example, in scenario 3C, difference based reasoners would assume that the final outcome (75 m^2 coverage with algae) can be perfectly apportioned into $25m^2$ that were caused by background causes and 50 m^2 attributable to the candidate cause, but would not consider that some portion of the $75m^2$ coverage is due to the influence of both the background and the candidate.

Similarly, in the multiplication strategy, because reasoners assume an interaction between background causes and candidate causes, they also violate the assumption of independence that a proportion strategy requires. Multiplication reasoners consider the candidate cause more as a moderator than an independent cause. Consider this example involving also mineral C and algae growth. According to the multiplication strategy, instead of looking at the explicit influence of mineral C towards algae growth (how efficient is mineral C in producing algae growth?), they are concerned with the implicit influence of mineral C (how efficient is mineral C in interacting with the background causes in producing algae growth?). Therefore, they considered the efficacy of mineral C as how many times more the background causes produce the growth when interacting with mineral C relative to without interaction.

Finally, besides these three strategies, the fourth possible approach for participants to make judgment would be to simply neglect the base rates (i.e. not considering the information of effect magnitude when the cause is absent). While this is not a reasoning framework *per se*, participants did have opportunity to opt for this, hence I included in the consideration during data analysis. Because this approach was the simplest to do, it was possible for participants to opt for this, especially when they were confused, or it was too difficult to adopt other strategies. For

---

would be to raise it to 1.

instance, to answer the question posted in scenario 5, participants who neglected base rate might simply responded that the area would be 75 m^2 because in scenario 3C, Q(e|c) = 75 m^2 after receiving treatment with mineral C, ignoring the specified 10 m^2 base rate.

1.3 Reasoning Situations

A potential factor of reasoning strategy is the situation in which the information for reasoning comes from. In the previous scenario 3C, the situation was about one pool, and within this pool, the area of algae before treatment changed from 25 m^2 to 75 m^2 after the treatment with the mineral. In this situation, the magnitude change was presented within the same entity, but happened across different time frames, i.e. before-after treatment.

Another possible situation would involve two different pools. In this situation, the control pool that does not receive any treatment serves as a reference to the other pool that receives treatment with the mineral. In other words, scenario 3C could be setup to be about two pools, in which the area covered by algae in the control pool would be 25 m^2, while the area in the treated pool would be 75 m^2. It is important to consider this, because reasoners might consider the base rate information (i.e. effect magnitude in the absence of candidate cause) might be less credible in the within- situation rather than a between-entity situation because the base rate effect magnitude was observable only in the between-entity. Consequently, this could trigger uncertainty to the constancy of the influence of background causes on the effect, which could lead a reasoning strategy that considers the interaction between the candidate and background causes (i.e. multiplication) to be more prevalent in a within- rather than a between-entity situation.

1.4 Strength vs. Structure

Both ΔP-derived difference and Power PC-derived proportion strategies consider the *strength* of a causal relation as the basis of causal reasoning. Griffiths and Tenenbaum (2005), however, argue that causal inference is concerned with determining the *structure* that generates the evidence: Instead of reasoning how strongly a candidate cause produces an effect, people first consider whether or not a causal relation exists at all. Griffiths and Tenenbaum proposed a Causal Support model, based on causal graphical models, or causal Bayesian networks (Pearl, 2000), to assess the causal structure between a candidate cause and an effect. The causal graphical model denotes any causal relation using nodes and arrows, with the nodes representing variables and the arrows representing causal connections between the variables. Nodes that the arrows are pointing to are referred to as "children" of the "parent" nodes where the arrow originates. Because this theory is based on Bayesian method, one of the critics is that it assumes that people are capable to apply holistic induction approach when there is evidence that people are holding onto heuristic approach (for details, see Gigerenzer & Todd, 1999).

Representing the above examples in a causal graph involves three nodes: a mineral node (candidate cause), an algae node (effect), and a background causes node. The two competing graphs (hypotheses) in these examples would be a 'mineral is non-causal' graph, where only the background causes node is connected with the algae node and a 'mineral is causal' graph, where both the background causes and mineral nodes are linked with the algae node to represent the causal situation. The main idea of Causal Support is to measure the extent of data favouring the causal graph over the non-causal graph. Griffiths and Tenenbaum (2005) represent the index of this favouring using the Bayes factor, i.e. the log likelihood ratio of these two causal graphs: $\log \frac{p(data|graph_{causal})}{p(data|graph_{non-causal})}$. For binary causation, they specify the parameterization as noisy-OR

for generative scenarios and noisy-AND-NOT for preventive scenarios. For continuous causation, however, it is not clear what would be the most suitable parametrization. While Griffiths and Tenenbaum presented Poisson process parameterization in their paper, this was conducted to discuss the results of previous studies of using rates, i.e. the number of times a binary effect occurs in a continuous time interval (Anderson & Sheu, 1995; Wasserman, 1990). Even though in principle the idea of the structure model could be extended to apply the continuous outcome scenarios here, determining the appropriate parametrization (and thus also generating predictions), is beyond the scope of this study. Thus, I exclude the support model from my considerations.

1.5 Notation

I use the following notation when referring to any condition in all experiments of Chapter 2 and 3: condition [Q(e|c):Q(e|¬c)]. The first represent effect magnitude in the presence of the cause, while the second term represent effect magnitude in the absence of the cause. Chapter 4 employs a different notation scheme.

Chapter 2: Empirical Investigation via Explicit Judgment

In this chapter, I present my initial attempt on studying causal reasoning with continuous outcomes. All experiments in this chapter adopted the standard causal question. In other words, after providing participants with information on magnitude changes with respect to the presence and absence of a cause, participants were asked to provide a judgment about the extent to which they thought the cause influences the change of effect magnitude.

2.0 Experiment Overview

Two experiments in this chapter adopted the same basic design, and adapted the same 15 conditions from Buehner et al. (2003). Originally, each of these conditions contained two bits of information capturing a simple relation between a binary cause and a binary effect. In this study, the first bit of the information, which originally corresponded to the probability of the effect when the cause was present P(e|c), was directly mapped to a continuous quantity of the effect when the cause was present Q(e|c). Analogously, the second bit of the information [originally corresponding to the probability of the effect when the cause was absent P(e|~c)], was directly mapped to the quantity of the effect when the cause was absent Q(e|~c).

The selection of these conditions allowed for a rich investigation between difference and proportion strategies. Figure 2.1 displays predictions of the difference and proportion indices of these conditions plotted against the base rates. In these conditions, there are groups of conditions that share the same difference index. They are connected together with straight lines. To ease the discussion, I used a naming convention to reflect common difference indices: any condition with common difference indices also shares the same group name. The difference index of each group is written in subscript. For example, the non-contingent conditions were labelled $C_{0.00}$, whereas

conditions [1.00:0.75], [0.75:0.50], [0.50:0.25], [0.25:0.00] received label $C_{0.25}$. See Table 1.2 for a complete list of all conditions, organized by grouping.

Table 1.2: Conditions for the experiments

Condition Group	Conditions	
	Q(e|c)	Q(e|¬c)
$C_{0.00}$	1.00	1.00
	0.75	0.75
	0.50	0.50
	0.25	0.25
	0.00	0.00
$C_{0.25}$	1.00	0.75
	0.75	0.50
	0.50	0.25
	0.25	0.00
$C_{0.50}$	1.00	0.50
	0.75	0.25
	0.50	0.00
$C_{0.75}$	1.00	0.25
	0.75	0.00
$C_{1.00}$	1.00	0.00

Note: The conditions are displayed for generative experiments. For preventive experiments, the roles of Q(e|c) and Q(e|¬c) were reversed, such Q(e|c) ≤ Q(e|¬c) .

The most apparent between these two strategies was the trend of the predictions. When predictions of the proportion strategy that had the same values of difference indices (i.e. in the same group) were connected, they formed upward trends in the contingent conditions, and flat lines in the non-contingent conditions (see Figure 2.1 for trends of the predictions). These trends showed the influence of base rates on the proportion strategy: for contingent cases in the generative scenario, as the base rate increases, proportion indices also increase, whereas in the preventive scenario, as the base rate increases, the indices decrease. In contrast, such influence of

base rates is not evident for the difference strategy; hence, they formed flat lines in all conditions for the difference strategy predictions.

Participants began by reading a cover story providing background and context information about the causal relation between the candidate cause and effect. Because the story accommodated all 15 conditions, it included clarification of variation between each condition. For each condition, participants received visual information of Q(e|c) and Q(e|¬c), and proceeded to the task. Participants were asked to give a strength rating for the cause in producing the effect by selecting a number on a scale of 0 to 10; with 0 indicating that the cause was absolutely ineffective in producing the effect, whereas 10 was otherwise.

Besides qualitative inspection of trends in the data, I also adopted a statistical hypothesis testing method to the trend. Specifically, depending on the normality of distribution, I compared the means across the same group using ANOVA or Friedman's Test procedure. Results from this test would reveal whether the trend was linear or otherwise. Linear trends indicated that participants response were independent of base rate, which the difference strategy predicts.

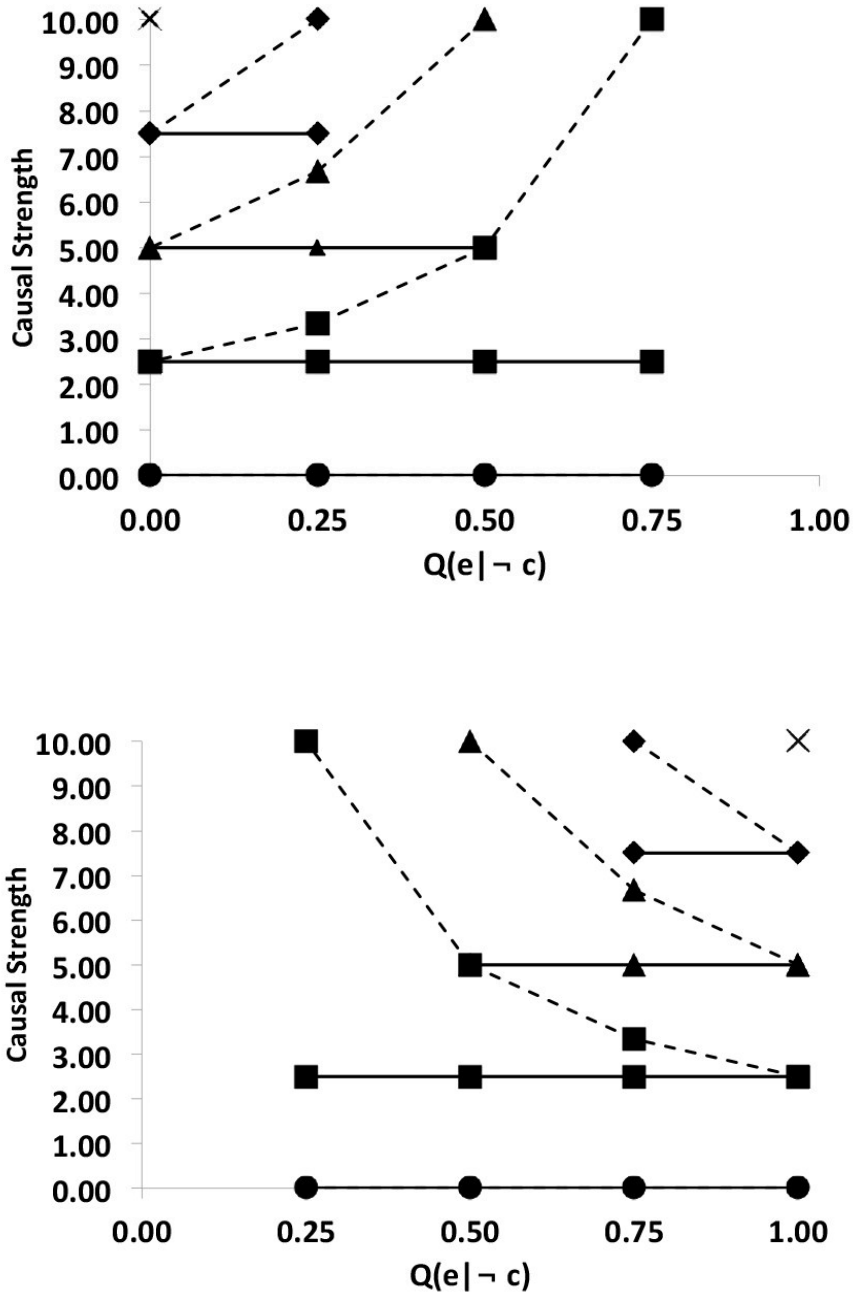


Figure 2.1: Plots of Predictions. Plots on the top and bottom correspond to the generative and preventive scenario, respectively. The dotted lines in the plots indicate predictions for the proportion strategy while the solid lines indicate predictions for the difference strategy.

Experiment 2.1

This first experiment aimed to compare judgments between directly mapped causal relations with their corresponding binary relations. Because of this, the experiments entailed four sections: generative-continuous, generative-binary, preventive-continuous, and preventive-binary. Participants' assignment to each section was determined randomly.

2.1.1 Method

Participants. 137 participants were recruited from a crowdsourcing platform, Maximiles. After completion, they received points which they could use for making purchases. Participation was limited to those whose English is their first language.

Design and Procedure. Participants accessed this experiment via Internet. Once they provided consent on the first page of the experiment webpage, the server randomly assigned them to either one of the sections of this experiment: generative-continuous, generative-binary, preventive-continuous, and preventive-binary. They completed only one of these sections. The procedure of each section was the same, but with a different cover story (see Appendix A for complete cover stories). In every section, participants completed 15 conditions.

For generative-continuous participants, they began by reading through a cover story about corn crops. The cover story told participants to imagine themselves as researchers of a study involving 15 different EU countries, each studied one fertilizer. The story continued by telling participants that in each country, there were two plots both with area of 100 meter square freshly sown with corn: the experimental plot received the fertilizer, whereas the control plot was left without any treatment. Participants were also told that at the end of the study, the area of usable corn crops were measured and recorded. After reading the story, participants received information for each condition in a visual format. The images consisted of two critical pieces of

information: the area of usable crop on the control plot, and on the experimental plot. The former information was on the right side, while the latter was on the left side of the display.

Participants in generative-binary also received the same cover story. Instead of two plots, however, there were two sets of 20 plots freshly sown with corn. One set was declared as control (i.e. receive no treatment), whereas the other set as experimental plots (treated with the fertilizer). After reading the story, similar to the generative-continuous, participants in this section also received information for each condition in the same visual format: the image on the left portrayed number of plots with crop, whereas on the right were plots without usable crop.

On the other hand, the cover story for participants in the preventive scenario was about chemotherapeutic agents in fighting tumours. To motivate the 15 conditions, participants were told that the study involved 15 laboratories, and each studied one chemotherapeutic agent. For continuous outcomes section, the story continued by telling participants that, also, in each lab were two mice with tumour grown in their brains. And the initial sizes of these tumours were 10 micrometres cubic. One of the mice received treatment with the chemotherapeutic agent, whereas the other did not (control). Participants were also made known that at the end of the study the sizes of tumours were measured and recorded. When answering the condition, participants saw images of mouse brains: on the left side was the image of brain with certain size of the tumour after treated with the chemical agent, whereas on the right was image of the control brain with its original tumour size.

For preventive-binary section, participants obtained the similar cover story with modifications: in particular the story explained that in each lab were two sets of 20 mice of which all of them had tumours. The first set was exposed to treatment whereas the other set remained untreated as control. After reading the story, similar to other sections, participants also

obtained the condition information in a visual format except the images were about 20 brains, some of which remained infected, whereas some were healthy.

In all sections, underneath those images was the instruction: "*Please provide a rating of the effectivity of this fertilizer [chemotherapeutic agent]*." To make the rating, participants needed to choose a number from a range of 0 to 10, with 0 indicated "*absolutely ineffective*" and 10 indicated "*absolutely effective*". After that, participant went through the next condition. Each participant received different computer-generated random sequence of conditions.

2.1.2 Results

Histograms of the judgments revealed non-normal distributions for all sections. Because of this, I plotted the medians of each condition with respect to base rates. A qualitative observation of these plots revealed that for the generative-continuous section, the trend was neither close to the difference nor the proportion predictions (see Figure 2.1), particularly at condition groups $C_{0.25}$ and $C_{0.50}$, whereas trends for other conditions leaned towards the proportion strategy. For other sections (generative-binary, preventive-continuous, and preventive-binary), the trends, qualitatively, followed prediction of the proportion strategy.

Applying Friedman's Test on the judgments revealed results that were in agreement with the qualitative observation for all sections except generative-continuous section. This is because Friedman's Test on judgments for this section produced non-significant results for all group conditions: $C_{0.00}$ ($\chi^2(3) = 2.586$, $p = .460$; $C_{0.25}$: $\chi^2(3) = 1.327$, $p = .723$; $C_{0.50}$: $\chi^2(2) = 2.116$, $p = .347$; and $C_{0.75}$: $\chi^2(1) = .167$, $p = .683$. This non-significant trend indicated that the majority of judgments were independent of base rate influence, which is also the prediction of the difference strategy.

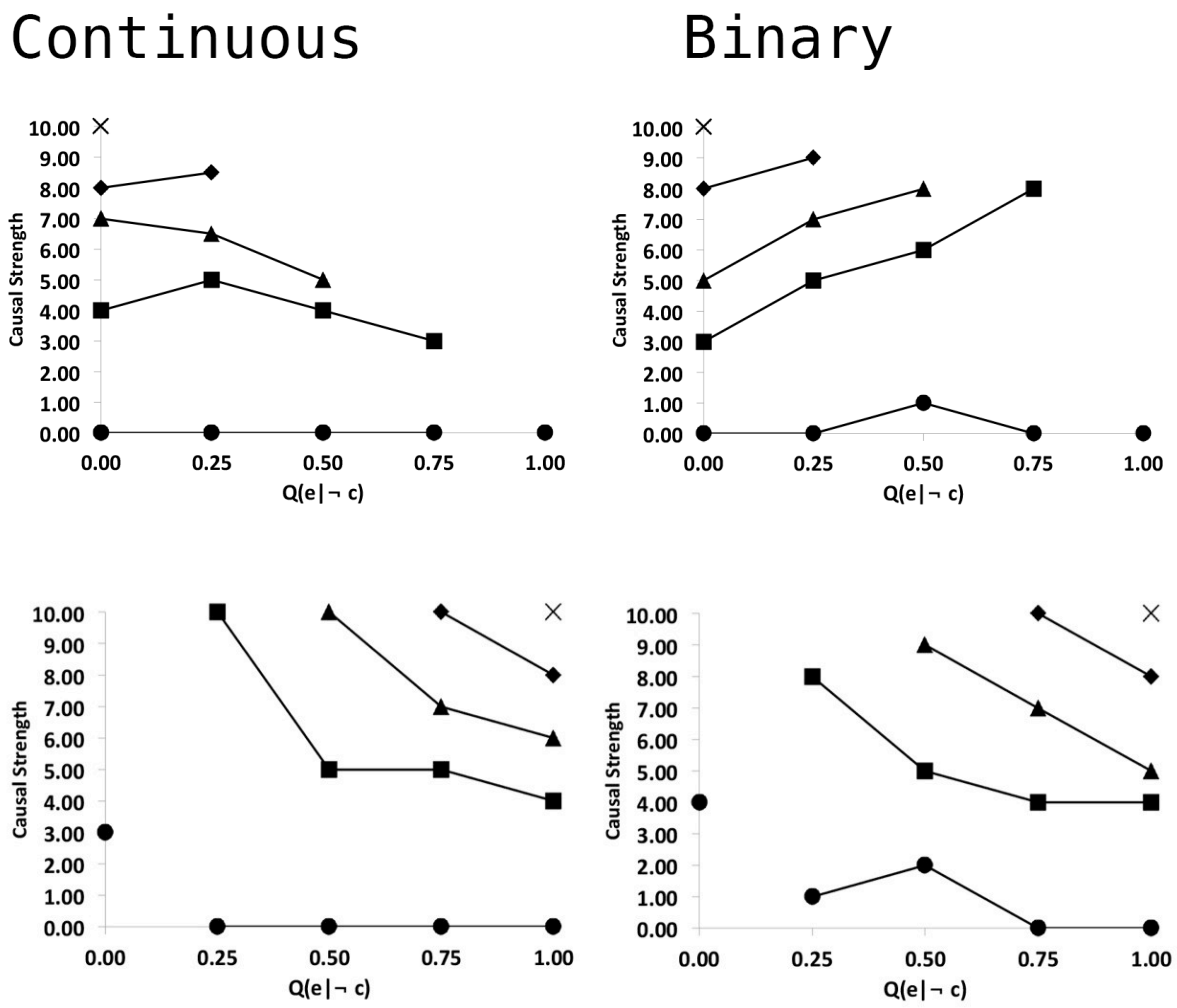


Figure 2.1: Results for Experiment 2.1. The left plots correspond to continuous scenario, while the right plots refer to binary scenario. The top and bottom row respectively represents results of generative and preventive.

For the generative-binary section, on the other hand, the test revealed that all group conditions had significant trends: $C_{0.00}$ ($\chi^2(3) = 16.121$, $p = .001$; $C_{0.25}$: $\chi^2(3) = 1.416$, $p \leq .001$; $C_{0.50}$: $\chi^2(2) = 17.487$, $p \leq .001$; and $C_{0.75}$: $\chi^2(1) = 18.615$, $p \leq .001$. The significant trend for condition group $C_{0.00}$ was probably contributed by the non-normative medians of condition [0.50:0.50].

For the preventive-continuous section, the results were non-significant for group $C_{0.00}$ (($\chi^2(3) = .698$, $p = .874$) but significant for other groups: $C_{0.25}$: $\chi^2(3) = 33.915$, $p \leq .001$; $C_{0.50}$:

$\chi^2(2) = 24.844$, $p \leq .001$; and $C_{0.75}$: $\chi^2(1) = 23.516$, $p \leq .001$. Relative to the other sections, results of this section were the closet towards matching with prediction of the proportion strategy.

For the preventive-binary section, the results were similar to its continuous counterpart except that results for group $C_{0.00}$ were non-normatively significant (($\chi^2(3) = 14.532$, $p = .002$) like the results of other groups: $C_{0.25}$: $\chi^2(3) = 35.693$, $p \leq .001$; $C_{0.50}$: $\chi^2(2) = 31.565$, $p \leq .001$; and $C_{0.75}$: $\chi^2(1) = 20.571$, $p \leq .001$. The non-normative trend for group $C_{0.00}$ was probably due to conditions [0.25:0.25] and [0.50:0.50].

2.1.3 Discussion

With respect to the significant results of groups $C_{0.00}$ in the generative-binary and preventive-binary sections, while this seems to deviate from predictions of the proportion strategy, further scrutiny on distributions of these conditions revealed that the modal response was actually zero. Because of a minority of non-zero responses, and Friedman's Test's negligence of any overlap, the results for this condition were significant.

Unlike in the preventive scenario where judgments of binary and continuous causation were in agreement (suggesting that the proportion was the dominant strategy,) in the generative scenario, only the binary section showed support for the proportion strategy. Most participants in the generative-continuous section opted for the difference strategy.

One possible explanation for this discrepancy would be the issue of limit saliency. Because the maximum continuous magnitude of an effect is context dependent, it is non-natural, and maybe less intuitive. In this experiment, while the story mentioned that a fertilizer can generate growth to cover up to 100 meter square, which was the artificial limit, participants could think that its maximum efficacy on growth could be more, if it was given more opportunity with a bigger plot area, for instance. Because the proportion strategy was sensitive to the upper

limit, perhaps for participants reasoning using this strategy was less instinctive when the limit was non-natural. Consequently, they were more comfortable to adopt limit-independent strategy, the difference.

This issue of limit is only applicable in generative-continuous section. Its preventive counterpart, however, does not have this issue because of its inhibiting nature. Even though in the preventive-continuous section the reasoning involves a deterministic entity as well, the maximum efficacy with which a cause can inhibit the magnitude of the effect would always be to zero. Regardless of any contexts, no continuous magnitude of effect can be reduced to be less than zero. For example, in the tumour situation, a cause can maximally impede the tumour until there was no more tumour, i.e. a magnitude of 0 micrometre cubic. Because of this, the results of the preventive-continuous mostly followed the proportion strategy. Experiment 2.2 continued the investigation by manipulating the saliency of the limit, specifically in generative scenario.

Experiment 2.2

This experiment aimed to examine whether saliency of limit influenced the judgment in generative scenario. To do this, I prepared two similar cover stories as the main manipulation of this experiment: one with a salient limit, whereas the other was not.

2.2.1 Method

Participants. 75 undergraduate students from the School of Psychology, Cardiff University participated to fulfil part of a course requirement. They had not been involved in any other experiments that were using the same cover story.

Design and Procedure. Participants were randomly assigned into three sections: clear-limit, no-limit, and binary. For all sections, participants received the same 15 conditions but different version of cover story. The first cover story was about a scenario without clear limit, the

second story put more emphasis on the limit and the third story was their binary counterpart as comparison. The general theme of these stories was about cloud seeding procedures. The stories explained that silver iodide was sprayed into clouds to initiate a merging process of tiny water drops, resulting in bigger and heavier drops that fall down as rainfall. The stories then described a project to study 15 chemical agents as alternative for silver iodide in the procedure. Involving 15 countries, the cover stories further explained that each country studied only one chemical agent using two locations: at one location, researchers sprayed the clouds with the chemical agent, whereas the other location was reserved as control without any treatment.

In the first cover story, the story continued by informing participants that the amount of rainfall at both locations was measured and recorded. In the second story, it additionally continued to inform participants that the relative humidity inside each of the areas was measured. To highlight the saliency of the limit, the story clarified that the maximum possible relative humidity is 100%, which means that humidity is so high that water will condensate and fall down as rain. In the third story, participants read that the study involved a selection of 20 cumulus clouds at the experimental location, which received spray of the chemicals. Further, at the control location, another 20 clouds were observed as control. The story further explained that researchers in the study observed and recorded any cloud that produced rainfall in these two locations. For complete cover stories in this experiment, see Appendix A.

In each section, after reading the cover story, participants received the first randomly presented condition. Information in each condition was presented in visual format consisting of images to represent effect magnitudes for base rate and post-treatment. Underneath these images, participants received the following instruction: "*Please provide a rating of the effectivity of this*

chemical agent." To make the rating, participants had to chose a number from 0 to 10, with 0 indicated the least strength.

2.2.2 Results

Qualitative observation on median plots for both limit sections revealed a similar trend between the two (see Figure 2.2): They both followed the difference strategy. The plot for the Binary condition, in contrast, showed more tendencies towards the proportion strategy. Trend analysis on these results further supported this argument.

For the clear-limit section, Friedman's Test produced non-significant results for all groups: $C_{0.00}$: $\chi^2(3) = .528$, $p = .676$; $C_{0.25}$: $\chi^2(3) = .888$, $p = .828$; $C_{0.50}$: $\chi^2(2) = 1.385$, $p = .500$; and $C_{0.75}$: $\chi^2(1) = .500$, $p = .480$. Results for the no-limit section were also similar where all groups were non-significant, $C_{0.00}$: $\chi^2(3) = 7.382$, $p = .061$; $C_{0.25}$: $\chi^2(3) = .835$, $p = .841$; and $C_{0.75}$: $\chi^2(1) = .818$, $p = .366$ except for group $C_{0.50}$: $\chi^2(2) = .800$, $p = .018$. Conducting the same test on judgments of the binary section produced non-significant results for groups $C_{0.00}$: $\chi^2(3) = 6.881$, $p = .076$; and $C_{0.25}$: $\chi^2(3) = 7.214$, $p \leq .065$; but significant results for other groups: $C_{0.50}$: $\chi^2(2) = 10.415$, $p = .005$; and $C_{0.75}$: $\chi^2(1) = 9.000$, $p = .003$.

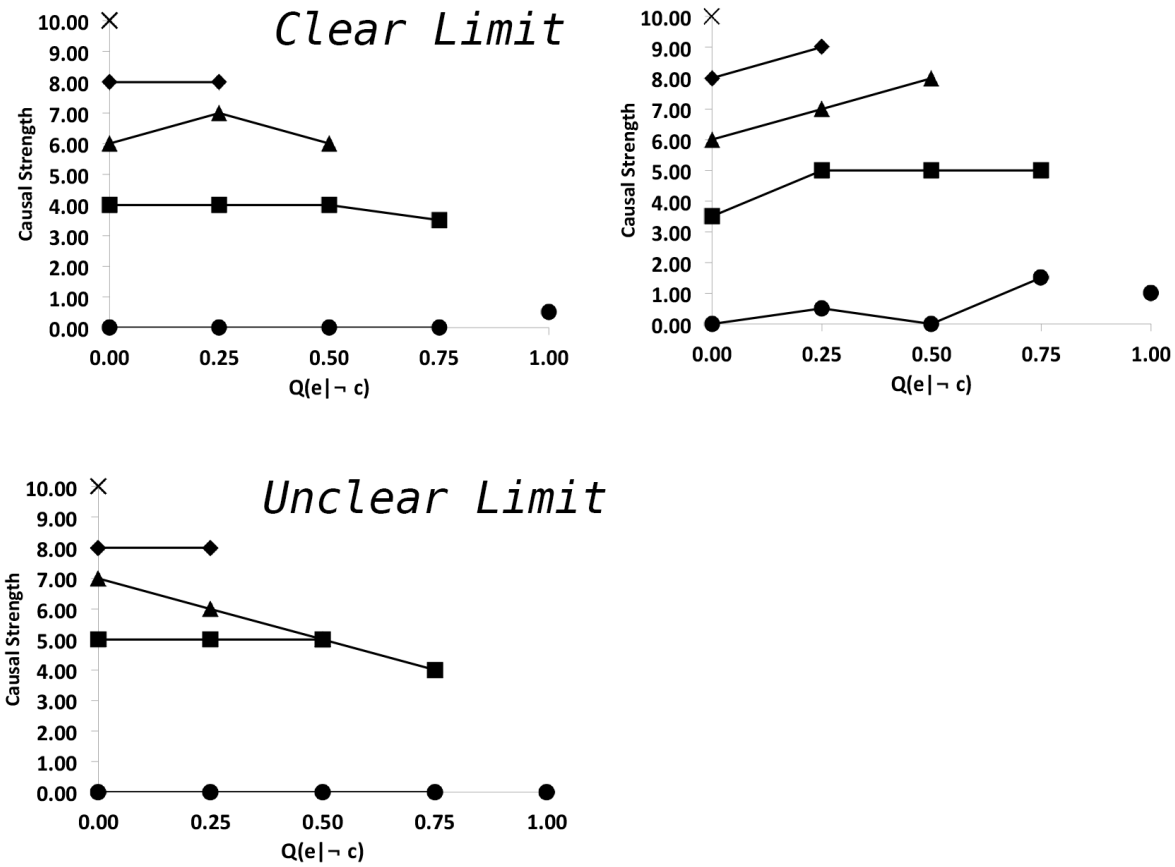


Figure 2.2: Results for Experiment 2.2. The first row contains results for two sections of the experiment: the left plot corresponds to continuous-clear limit section, whereas the right corresponds to binary section. Underneath these plots are the results for continuous-unclear limit.

2.2.3 Discussion

Contrasting results of clear-limit and no-limit revealed similar trends except for group condition $C_{0.50}$ in the no-limit section. While the trend for this group went up as the base-rates increased (despite not significant) in the clear-limit, it significantly went down in the no-limit condition. Inspection of histograms for condition [0.50:0.00] in both cases exhibited the same modal response of 5 but it was noisier in the no-limit section resulting to higher median of 7.

As for condition [0.75:0.25] in both limit sections, their histogram revealed a bimodal distribution where responses in accordance to the difference and proportion strategy were competing, surprisingly even in the no-limit section. Similar inspection for condition [1.00:0.50] exposed a more consistent modal response of 5 in both limit sections. In general, the judgments in both limit conditions were similar, suggesting that limit saliency might not be the major influencing factor.

As for the binary section, the results did not exactly follow predictions of the proportion strategy, as in Experiment 2.1, especially the non-significant trend of $C_{0.25}$. Further scrutiny on the distributions of conditions in this group revealed that the critical conditions were [1.00:0.75], [0.75:0.50] and [0.50:0.25]: For conditions [1.00:0.75] and [0.75:0.50], their distribution was bimodal indicating a competition between the difference and proportion strategy. Further, distribution for condition [0.50:0.25] was more complex because its modal response of 5 matched with neither predictions of the difference nor proportion.

Extended scrutiny of condition [0.50:0.25] in the clear-limit and no-limit sections, surprisingly, revealed similar phenomena to the binary section:the majority of judgments were 5, which matched with neither the difference nor proportion strategies. One potential explanation for this would be that participants might have thought that the chemicals and background causes

mutually-exclusively influenced the amount of rainfall. Thus, for this condition, they might consider the chemicals producing 25 units of rainfall, and background causes producing another 25 units of rainfall. In other words, the candidate cause produced half of the rainfall, while background causes produced another half. Therefore, when asked to give a rating of causal strength of the candidate cause, participants might have considered 5, which was the half of 10 as the maximum possible effective rating of the scale.

2.3 General Discussion of Chapter 2

The first experiment in this chapter showcased a comparison of directly mapped continuous relations with their corresponding binary relations. The results revealed that the proportion strategy was dominating in both scenarios of binary causation, but dominated only in the preventive scenario of continuous causation. In the generative continuous section, in contrast, the majority of participants reasoned using the difference strategy. This signals asymmetry between the generative and preventive scenario when involving causal relations with continuous outcomes. Because the most apparent distinction between generative and preventive continuous relation was the issue of limit saliency, the second experiment focused on this issue.

In the second experiment, I compared between-participants judgments between continuous causal relations with clear limit, and a relation with no limit. Also, I included their binary correspondence. In general, participants from both limit conditions responded to the task similarly. Two possible explanations for this were that there was really no influence of limit saliency, or the method of study did not successfully capture the influence. While the first argument is still open for discussion, the second argument already has support from the results of the binary section in Experiment 2.2. This is because, if the study method (i.e. a question asking participants to report causal strength) was sensitive, I expected that the proportion strategy

should dominate the results, in line with the results of the binary generative section of Experiment 2.1. The inconsistency of results of the binary generative sections in Experiment 2.1 and 2.2 signals an important issue with the current method (See Figure 2.3 for compilation of predictions and results for both experiments). Further, Buehner et al. (2003) have already identified the validity issue of this method.

Specifically, the rating scale in these experiments and others (binary causation) might be ambiguous. The ambiguity might have led participants to be confused between reliability and strength, resulting in conflation of these in their judgments. Moreover, Buehner et al. argued that the question asking for strength of a candidate cause is ambiguous with respect to the context of which the question appears to: i.e. whether the question was asking in the current learning context, or in a "counterfactual" context where alternative causes were absent.

Buehner and colleagues argued that if participants interpreted the question in the experiments as "*What difference does the candidate cause make in the current learning context where background causes already produce effect in a certain magnitude of the entity?*", they would response in accordance to the difference strategy; whereas, if participants interpreted the question as "*What difference does the candidate cause make in a new context when background causes never produce effect?*", their judgments would correspond to the proportion strategy.

Hence, in Chapter 3, I continued the investigation by adopting this new approach. More specifically, in the empirical work of that chapter, instead of asking participants to report vague rating about strength of candidate cause, I asked them to produce estimate of effect magnitude a candidate cause could produce in a novel "what if …" scenario.

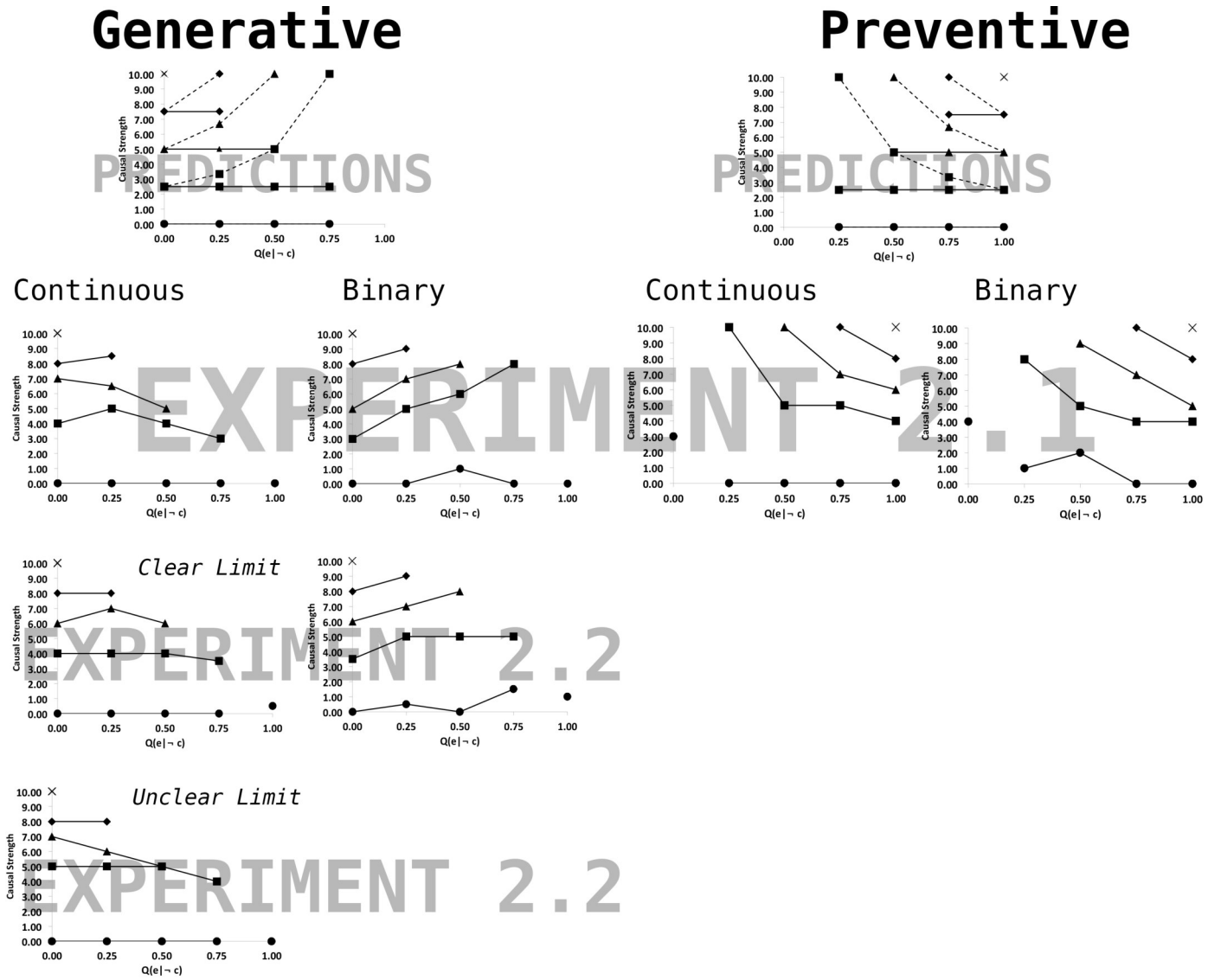


Figure 2.3: Predictions and Results for experiments in Chapter 2. The first row contains prediction plots. The left column plots corresponds to generative scenario, whereas the right column plots correspond to preventive scenario.

Chapter 3: Empirical Investigation via Hypothetical Judgment

Research in binary causation has shown that intervention (versus observation) elicits better causal understanding between candidate cause and effect because intervention warrants only forward inference, whereas observation warrants both forward and backward inference (Sloman & Lagnado, 2005). A classic approach to elaborate on this would be to consider an example of a causal relation between smoking and yellow teeth. If a reasoner intervenes by smoking, then he or she should predict an increase of probability that his or her teeth would become yellow (i.e. forward inference is warranted). In contrast, if the reasoner whitens his or her teeth, this intervention has no influence over his or her smoking habit (backward inference is not warranted). This is because, when a reasoner intervenes, he or she is "undoing" the link of the effect with its normal cause. In contrast, observation has no such effect. Observing somebody with white teeth would suggest that he or she is less likely to be a smoker. In this case, backward inference does also make sense.

Meder, Hagmayer, and Waldmann (2009) provided evidence that people are capable to derive interventional predictions from the unseen actions of observational knowledge. In the beginning of the article, they also discussed the difference between hypothetical intervention and counterfactual intervention: "The crucial difference between modeling hypothetical and counterfactual actions is that the latter require us to take into account the diagnostic information provided by the factual observation." In other words, counterfactual intervention combines inferences from both observation and intervention. While the content presented in this chapter, arguably, might be more appropriately referred to as hypothetical, I used the term "counterfactual" throughout to be consistent with the literature, i.e. Buehner et al. (2003) who initially coined this term for this experimental approach. In addition, the study of Meder,

Hagmayer, and Waldmann indicated that people did not adequately distinguish between hypothetical vs. counterfactual intervention.

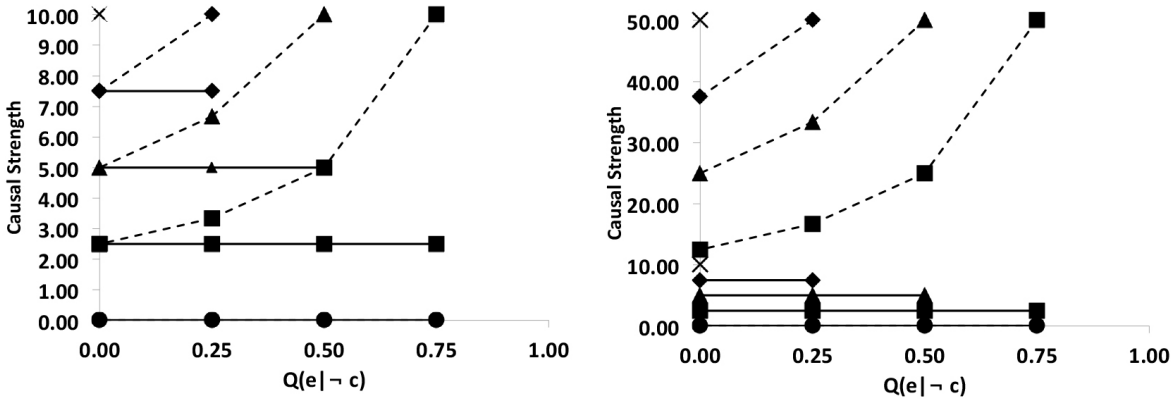
The counterfactual judgment procedure in this chapter entails two stages: learning of the causal strength, and applying this knowledge to a novel counterfactual scenario. In the first stage, the reasoner acquires information through the observable evidence and forms an inference of the candidate cause's strength in producing the effect via a specific reasoning strategy. When given a counterfactual scenario in which the candidate cause is acting on a new effect, the reasoner applies the learned causal knowledge to the counterfactual scenario to make a judgement about the (hypothetically) observed effect. One underlying assumption in this procedure is that the reasoning strategies in the learning stage and in the application stage are the same. Given that the general contexts and their scenarios between these two stages are so similar, it is reasonable to assume that people consistently use the same strategy in both stages. This chapter reports ten experiments, six of which concerned generative, and four preventive scenarios.

3.0 Experiment Overview

Similar to experiments in Chapter 2, all experiments in this chapter followed the same design involving the same 15 conditions, adapted from Buehner et al. (2003). For discussion in this chapter, I will also use the same condition naming convention as in Chapter 2 (see Table 1.2 for details). Thus, the predictions are also the same. See the first row in Figure 3.1 for the plot of the predictions for both the proportion and difference strategies.

Participants began the experiment by reading a cover story that provided context and background information about the causal relation between the candidate cause and effect. This information reflected the reasoning situation for that particular experiment, either a within- or a between-entity situation. In a within-entity situation, the story included a setting where the

changes of the effect magnitude in the presence and absence of candidate cause happened in the same pool, whereas in a between-entity situation, the changes of effect magnitude were between two different pools.[2] Because the story covered all 15 conditions, it also provided motivations for variations of base rates across the conditions. In the mineral-algae story example, different microclimates at different laboratories justified the differences between base-rates between the conditions. In generative scenarios, participants also received a rationale for the artificial limit. For the mineral-algae example, because the magnitude of algae growth was measured in terms of the area of water surface of the pool covered by algae, the size of pool surfaces served as the artificial limit in the example. The story ended with an overview of the task that participants needed to do (see Appendix A for the full list of cover stories for all experiments).



[2] see Chapter 1 for details between these two reasoning situations (within- vs. between-entity)

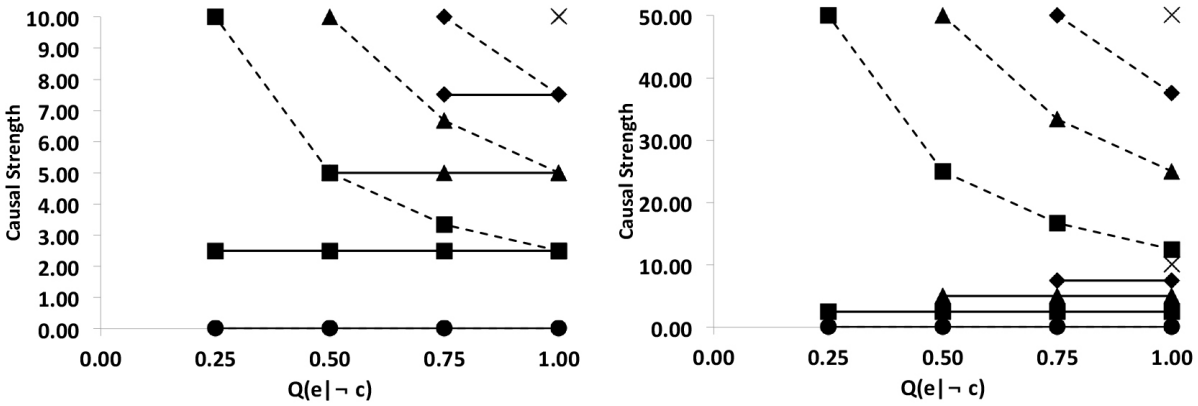


Figure3.1: Plots of Predictions. Plots on the top and bottom correspond to the generative and preventive scenario, respectively. The left and right columns contain the plots for the consistent limit and higher limit, respectively. The dotted lines in the plots indicate predictions for the proportion strategy while the solid lines indicate predictions for the difference strategy.

For each condition, participants obtained information about effect magnitudes with and without the candidate cause. In the above example, this information corresponds to the initial algae coverage of 25 m^2, and its increment to 75 m^2 after receiving treatment of mineral C. Then, participants received a question asking them to imagine a new entity with the same artificial limit as in the story (in the generative scenario only), i.e. a new pool with 100 m^2 surface area. The question continued by mentioning a specific base rate (depending on conditions), and asking participants to specify the magnitude of the effect if the candidate cause presented in that condition were applied to this new situation. For example, participants could be told that 10 m^2, of the 100 m^2 area of the new pool was already covered with algae, and then be asked to compute the area of algae covering the pool if the same mineral C were administered to this pool. After the question, participants received a second question. The second question replicated the first question but with a higher artificial limit than mentioned in the cover story, e.g. 500 m^2. The use of two limits in this procedure afforded a deeper investigation of the reasoning strategy, as explained below.

The example beginning of this chapter demonstrated an upper limit for causal influence that was *consistent* with the learning scenario (i.e. the size of the pool in the counterfactual judgment scenario was identical to that used in the learning situation). The second question for this example would be: How big would the area covered by algae be, if the same mineral C from scenario 3C in Chapter 1 were administered into a new pool with 500 m^2 surface area? In this question, the new limit was set higher to 500 m^2. For a reasoner who strictly adopts the difference strategy, this *higher* limit should make no change to his or her judgment relative to the original limit (i.e. with efficacy of 50 m^2, given the initial area was already 10 m^2, the judgment should also be 60 m^2). On the other hand, a proportion-based reasoner should respond to this new limit with different judgments relative to the original limit. He or she should first compute a new potential maximum efficacy mineral C could have in this scenario as 490 m^2 (500 m^2 – 10 m^2). The final judgment would then be an addition of 326.63 m^2 (i.e. 66.6% efficacy of mineral C in scenario 3C from the new potential maximum efficacy of 490 m^2) to the already 10 m^2 (326.63 + 10 = 336.63 m^2). From this example, it is clear that proportion-based judgments should be influenced by variations in limit, whereas difference-based judgments should not. Thus, having two limits provided more room to further probe the strategies participants adopted.

I pursued three analyses on the judgments to identify which strategy (proportion or difference) was dominant. The first and third employed a statistical hypothesis testing approach, while the second analysis involved qualitative inspection of the results.

3.0.1 Trend Analysis

This was the ubiquitous approach in the binary causation literature. The key to this conventional approach was the checking for trends in the data, particularly in conditions from the same groups, to see which trends (if any) from the predictions (see Figure 3.1) manifested in the

data. Because this approach relied on a central tendency measure, it is predicated on the assumption that there is (only) one dominating strategy. If there are competing strategies, and participants do not consistently apply them between conditions, or some participants adopt one, and others the other strategy, then the distribution of data may be bi-modal or even multi-modal; hence any approach based on a measure of central tendency will be compromised. Relying only on trends across all conditions for a verdict a dominant reasoning strategy could be problematic in this case.

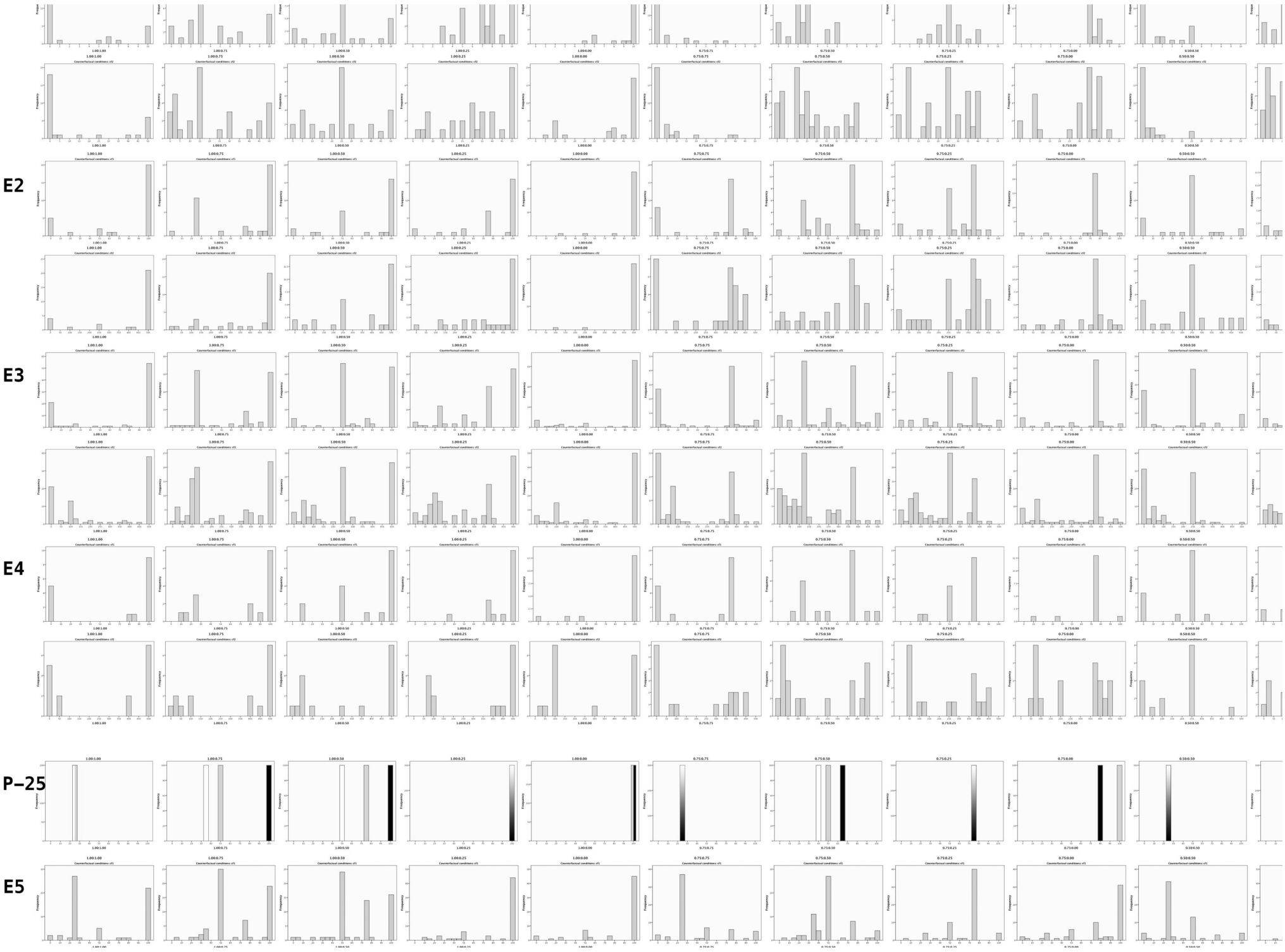
3.0.2 Histogram Analysis

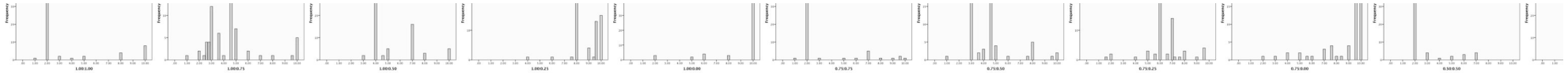
While I still used trend analysis on the data, mainly for the purpose of comparison with the literature, I also adopted another approach to supplement it. The main idea of this analysis was still to compare the judgments to the predictions, but rather than focussing only on central tendency, this considered the distribution of all judgments values. Figure 3.2 plots the predictions of each condition as expressed under this approach. In this figure, grey bars marked predictions for the difference strategy, black bars marked predictions for the proportion strategy, and white bars marked predictions for base rate neglect. In some conditions, all of the predictions converged (fully-overlapping, marked by gradient bars), whereas in other conditions, they were perfectly distinguishable (non-overlapping); there were also conditions where two predictions converged (partial-overlapping). Because this approach is based on histograms, determination of the bin sizes that define the histograms has to be sensible and systematic. For the 15 conditions of the experiments reported in this chapter, I have chosen the bin size to be 0.42, 2.08, 4.17, and 20.83 for experiments of which the range of judgments vary from from 0 to 10, to 50, to 100, and to 500, respectively. See Appendix B for discussion on selecting these bin sizes.

3.0.3 Tendency Analysis

Besides these, I adopted another approach to analyse each participant's tendency for adopting a particular strategy. More specifically, I used each participant's judgments to derive three scores, indicating tendencies to adopt the difference, the proportion strategy, or base rate neglect: Any judgment that 'matched' the prediction of a particular strategy contributed one mark towards the score of that strategy.[3] The average of each score computed over the 15 conditions represented the tendency of a participant to use that strategy (Note that for the proportion strategy, the tendency score was computed by averaging only over 14 conditions, because for the 15th condition, the prediction is undefined: i.e. for generative it was when $Q(e|\neg c) = 1.00$, while for preventive, it was when $Q(e|\neg c) = 0.00$). For the purpose of simplicity, I conducted this analysis only towards data of the consistent limit condition.

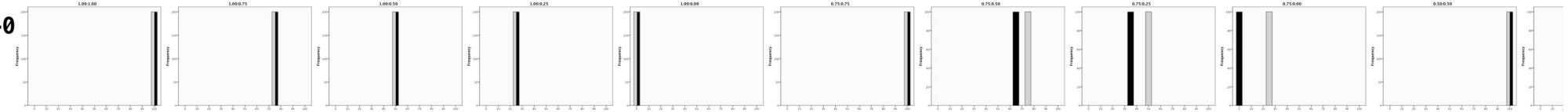
[3] For the purpose of this analysis, I defined 'matching' as being within ± 4.17% from the prediction. See Appendix B for discussion on this gamut definition.


E2
E3
E4
P–25
E5

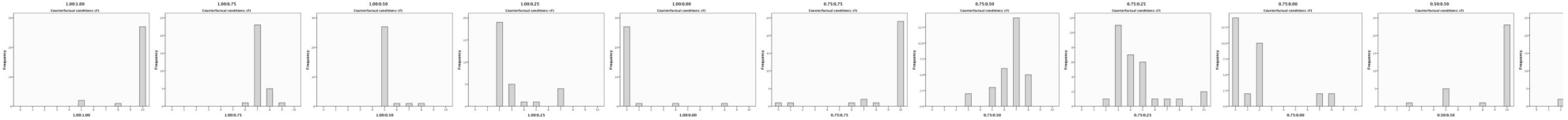


Preventive

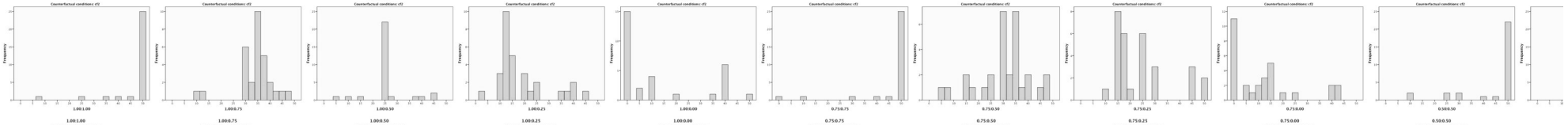
P-0



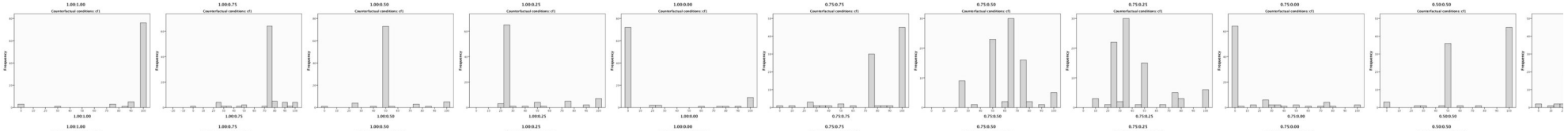
E7



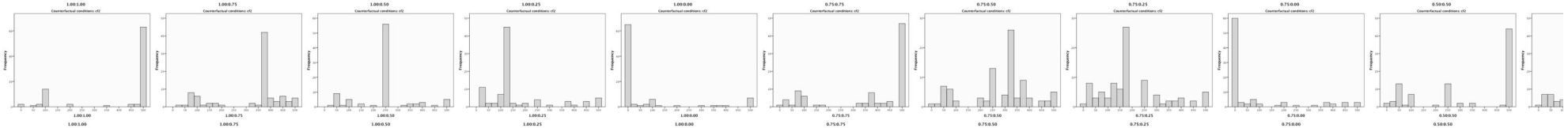
E8



E9



E10



Experiment 3.1

Experiment 3.1 examined reasoning about continuous outcomes in a generative scenario using the counterfactual judgment approach, as outlined above. In this experiment, the context was about skin rash as side effect of cosmetic creams.

3.1.1 Method

Participants. 30 undergraduate students from the School of Psychology, Cardiff University participated to fulfil part of a course requirement. Those who have participated in other experiments that used similar cover story as in this experiment were not allowed to participate.

Design and Procedure. Each participant worked on the 15 conditions listed in Table 1.2. Prior to the actual judgment task, participants were presented with a cover story about a study of skin rash as a side effect from the use of cosmetic creams. The story described 15 laboratories (corresponding to 15 conditions) that took part in the study; each studied only one cream on a patient. Participants were told that there were patients who already had rash even before the study and that at the beginning of each study, the area of skin rash on each patient's back prior to any application of the cream was measured, followed by the application of the cream to cover 10 centimetres squares of the skin area; an hour after this, the area of skin rash was re-measured.

After reading the story, participants went through each condition in the judgment task. The information pertaining to each condition, i.e. the magnitude of skin rash before and after treatment, presented visually using a silhouette of a man with a certain area marked as skin rash. This presentation involved three columns (see Figure 3.3 for a screen capture): the silhouette in the left column indicated the skin area before treatment, and was accompanied by the text "*The cream was applied on an area of 10 cm^2 on the patient's back. X cm^2 of this area was already*

covered by rash before the cream was applied.", whereas the right column portrayed the skin area after treatment, and was accompanied by the text "*After one hour, Y cm^2 of the 10 cm^2 area, where cream was applied, was covered by rash.*" In addition, a central column showed the legend of skin area when it was completely covered, or not covered at all with rash.

Underneath this visual presentation, participants received the following instruction: "*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 after one hour?*" Participants had to make the judgment by entering numbers from 0 to 10 in an empty box provided on the screen. After submitting their judgment, participants received another question with the same wordings except the area of 10 cm^2 was replaced with 50 cm^2. This time, participants had to type in their judgments from 0 to 50 in the empty box. Participants went through the same process in the same visual format across all conditions.

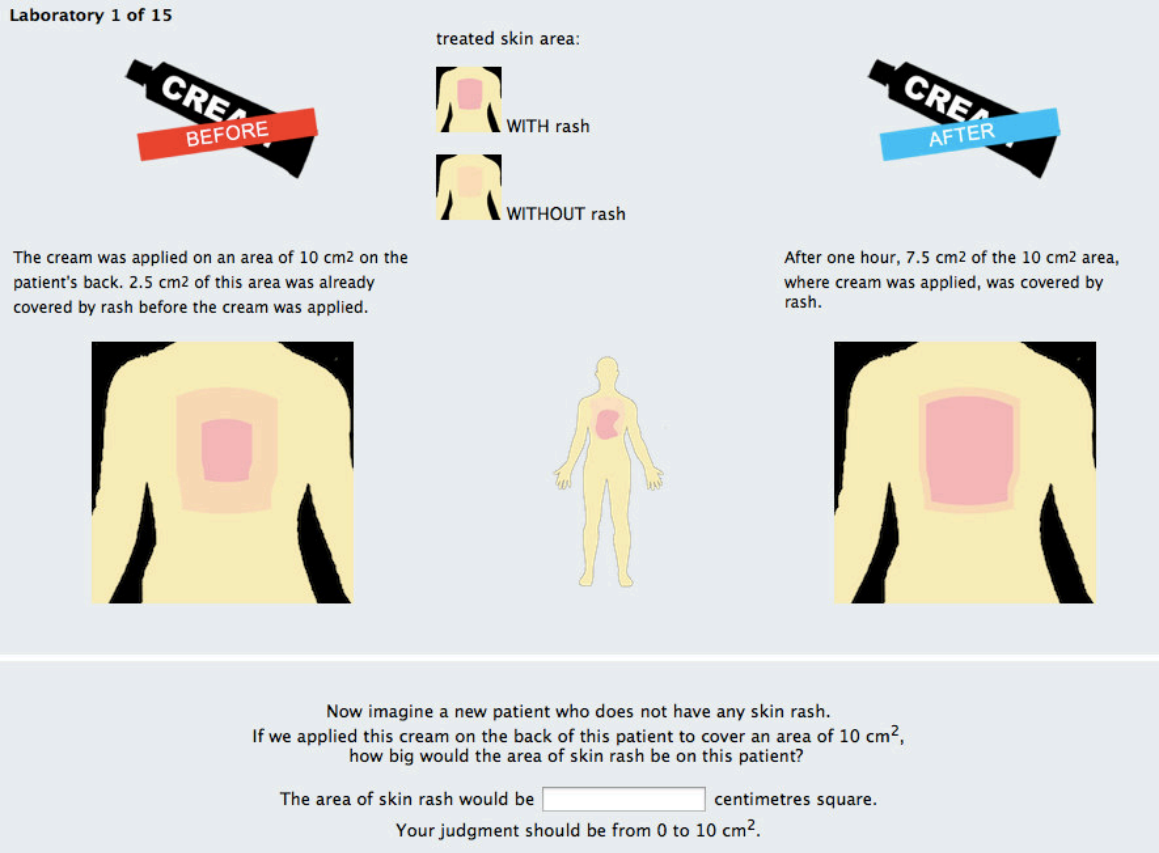


Figure 3.3: Screen capture of condition [0.75:0.25] in consistent limit condition for Experiment 3.1

3.1.2 Results

Figure 3.2 shows the distribution of judgments for each condition for both limits. Because some distributions were quite skewed, the trend plot analysis used the medians instead of the means. Comparing the trend plots of the medians (Figure 3.4) with the prediction plots (Figure 3.1), qualitatively, both the consistent and higher limit plots followed the flat lines trend. Friedman's test on the consistent limit judgments produced a significant result only for group $C_{0.00}$ ($\chi^2(3) = 14.679$, $p = .002$),[4] and non-significant results for other groups ($C_{0.25}$: $\chi^2(3) = 7.043$, $p = .071$; $C_{0.50}$: $\chi^2(2) = .228$, $p = .892$; and $C_{0.75}$: $\chi^2(1) = 1.190$, $p = .275$). For the higher limit condition, the same test also produced a significant result only on judgments for group $C_{0.25}$ ($\chi^2(3) = 12.991$, $p = .005$), and non-significant results on other groups ($C_{0.00}$: $\chi^2(3) = 6.300$, p

= .098; $C_{0.50}$: $\chi^2(2) = 1.357$, $p = .507$; and $C_{0.75}$: $\chi^2(1) = 1.087$, $p = .297$). The absence of slopes in conditions sharing the same base-rate suggests that participant's judgments were based on assessing the difference between Q(e|c) and Q(e|~c), rather than a proportion.

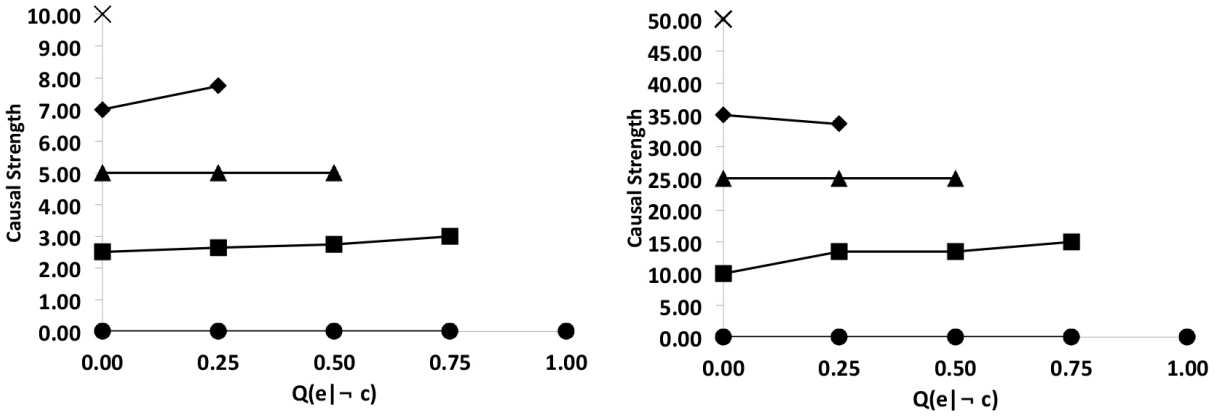


Figure 3.4: Results for Experiment 3.1. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

Results from the histogram analysis also agreed with the trend analysis. In all conditions, the highest number of judgments reflected a difference strategy, despite overlapping with the proportion strategy in eight of the conditions.

Results from the tendency analysis further supported this claim: Base rate neglect had the lowest mean score ($M = .24$, $SD = .02$), followed by the proportion strategy ($M = .40$, $SD = .03$). The highest mean score was for the difference strategy ($M = .69$, $SD = .05$). A Greenhouse-Geisser corrected ANOVA showed a significant effect of strategy $F(1.22,35.39) = 72.12$, $p \leq .001$. Bonferroni *post hoc* test further revealed that the score for base rate neglect was significantly lower than difference ($p \leq .001$) and proportion ($p \leq .001$) strategies. The difference between the difference and proportion strategy scores was also significant ($p \leq .001$).

[4] Despite being a flat line, this results is significant. I explain in the discussion below.

3.1.3 Discussion

In the trend analysis, any significant result of Friedman's test indicates that the conditions were all not the same; while the proportion strategy predicts that they should be different in the contingent conditions (because of the upward trend), the difference strategy would predict that they are all the same. In other words, the significant trend for group conditions $C_{0.00}$, would suggest that the results for conditions in this group were different despite being observed as a flat line.

Further inspection of the judgment distributions of this group (see histograms in Figure 3.2) revealed that (i) the modal responses were referring to zero; (ii) and there were minority judgments other than zero. Because Friedman's Test neglects judgments that were ties (i.e. the majority of zeros judgments), the significant results were driven by those minority judgments that exhibit an outcome density effect (Buehner et al., 2003). In binary causation, outcome density effect refers to judgments that follow the measurement of probability of effect being present regardless of the presence of candidate cause, i.e. $P(e)$.

In the higher limit scenario, although the lines were flat, suggesting use of the difference strategy, the judgments themselves deviated from the difference strategy prediction. If participants strictly made judgments in line with the difference strategy, the magnitudes of judgments in the higher limit scenario would range from 0 to 10, adhering to the *absolute* differences principle in the prediction. In contrast, the judgments were *scaled-up* to accommodate the new, higher limit. In short, participants did adopt a difference strategy in this experiment, as corroborated by all of the analyses, but they were also attentive towards the role of limit: They considered the influence of the limit by scaling their judgments using the same

scaling factor between the learned limit and the counterfactual limit (in this experiment, it was a factor of 5 to scale 10 cm^2 to 50 cm^2).

One concern in this experiment, however, was the plausibility of the cover story. From the story, participants were aware that the researcher in the story would initially measure the rash area, apply the cream, and re-measure the area where rash has broken out. This situation was reasonable in conditions with base rates of zero, but not in conditions with non-zero base rates. This is because, in these conditions, the story would then suggest that after the initial measurement of the rash area, the researcher would apply the cream even on top of skin that had already broken out with rash. While this make sense if the cream was to heal the rash, it does not when the idea of applying the cream was to study its side effect of skin rash, as in this story (for the complete text of the story, see Appendix A). Experiment 3.2 aimed to address this concern.

Experiment 3.2

Experiment 3.2 was a replication of Experiment 3.1 but using a different cover story to address the concern of the cover story in Experiment 3.1. Specifically, the cover story explained that after the initial measurement of skin rash area, the cream was applied even when the skin was already broken out with rash. This is implausible as the study mentioned in the story aimed to measure skin rash as a side effect of the cream. Thus, in this experiment, participants experienced a new cover story of fertilizers influencing crop growth in the same counterfactual judgment format as in the previous experiment.

3.2.1 Method

Participants. 30 undergraduate students from the School of Psychology, Cardiff University participated to fulfil part of a course requirement. The same exclusion criteria with Experiment 3.1 were adopted in this experiment.

Design and Procedure. This experiment used the same 15 conditions (see Table 1.2). The cover story was adapted from Experiment 2.1 in Chapter 2 (i.e. influence of fertilizers on crops growth), with some modifications at the end of the story to reflect the counterfactual task that participants needed to do. Specifically, the last paragraph was changed to this: "*For each fertilizer, we are asking you to consider how effective you think it is in promoting corn yield. To do so, we are asking you to imagine a new field of freshly sown corn that would show no yield in the absence of fertilizer. We are then asking you to imagine how much of that field would yield corn, once the fertilizer would be applied.*"

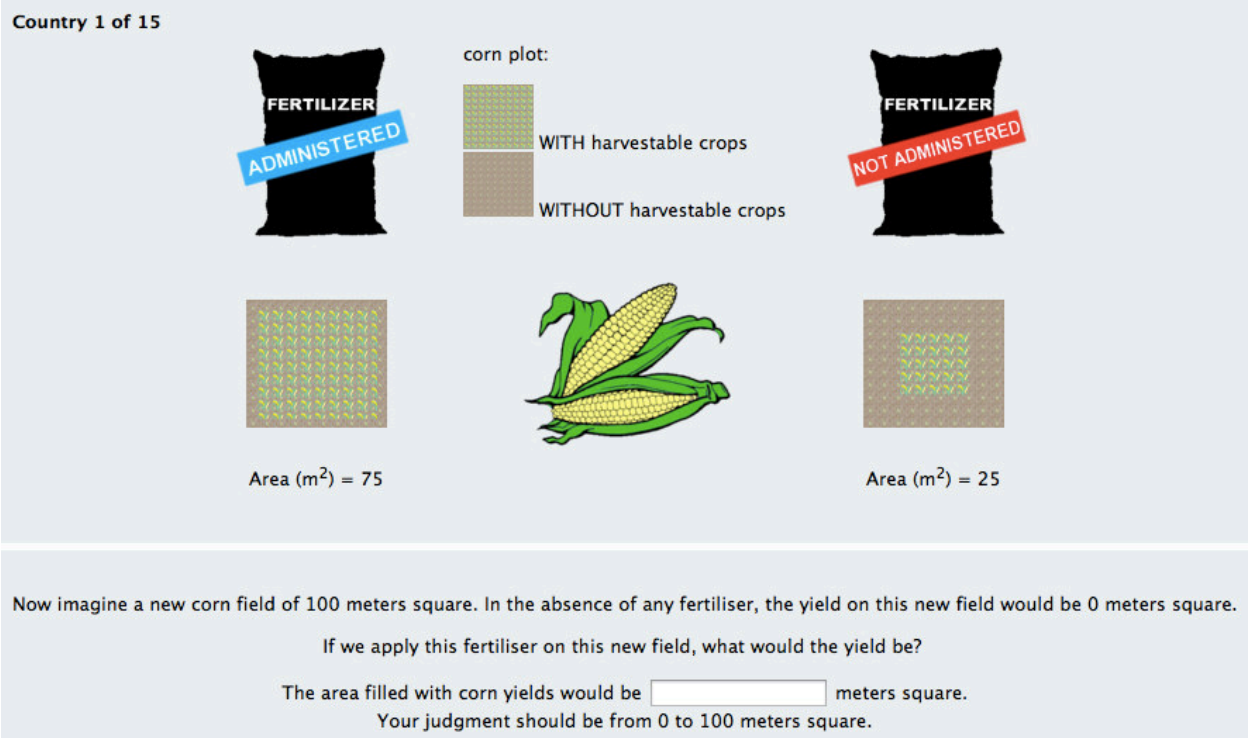


Figure 3.5: Screen capture of condition [0.75:0.25] in consistent limit condition for Experiment 3.2

Similar to Experiment 3.1 of this chapter, after reading the story, participants received information for each condition in a visual format, consisting of three columns as in the previous experiment. The left column portrayed the area of usable crops on the experimental plot, while the right column portrayed the area of usable crops on the control plot. The middle column

showed the legend of the area of a plot completely filled with harvestable corn crops, and of a completely barren plot, i.e. no crop growth (see Figure 3.5 for a screen shot). Underneath these images was the judgment question: "*Now imagine a new corn field of 100 meters square. In the absence of any fertiliser, the yield on this new field would be 0 meters square. If we apply this fertiliser on this new field, what would the yield be?*" Participants had to give judgment by filling in an empty box with numbers from 0 to 100. The higher limit question followed after that changing the limit of the corn plot to "…*a new corn field of 500 meters square*… ." After that, participant went through the next condition.

3.2.2 Results

Distribution of the results required consideration of the median for the trend analysis plot (see Figure 3.6). Qualitative comparison of these plots with the prediction plots in Figure 3.1 suggests that participants did not adhere to a single strategy: For the consistent limit, judgments for conditions $C_{0.50}$ and $C_{0.75}$ clearly reflected the difference strategy. Judgments for conditions $C_{0.25}$, indicate a noisier difference strategy because of its non-linear tendency at higher base rates. Conditions $C_{0.00}$, on the other hand, most closely reflected neglect of base rates. Friedman's test on these judgments supported this observation where all trends were non-significant (i.e. $\chi^2(3) = 7.043$, $p = .071$ for group $C_{0.25}$; $\chi^2(2) = .228$, $p = .892$ for group $C_{0.50}$; and $\chi^2(1) = 1.190$, $p = .275$ for group $C_{0.75}$) except for group $C_{0.00}$ ($\chi^2(3) = 14.679$, $p = .002$). For the higher limit, only judgments for conditions $C_{0.50}$ indicated a difference strategy. Conditions $C_{0.75}$, in contrast, did not fit in any predictions. The other conditions suggested judgments corresponding to base rate neglects. Friedman's test on these judgments produced all non-significant results (i.e. $\chi^2(3) = 6.300$, $p = .098$ for group $C_{0.00}$; $\chi^2(2) = 1.357$, $p = .507$ for group $C_{0.50}$; and $\chi^2(1) = 1.087$, $p = .297$ for group $C_{0.75}$) except for group $C_{0.25}$ ($\chi^2(3) = 12.991$, $p = .005$).

Histogram analysis on the data revealed that participants made judgments that seem to suggest base rate neglect. This was evident in histograms for all conditions, including in conditions $C_{0.50}$ (despite trend analysis suggestion of difference strategy).

Tendency analysis unveiled a significant main effect (Greenhouse-Geisser corrected) of the three scores, $F(1.12, 32.41) = 7.28$, $p = .01$. The mean score for base rate neglect was $M = .60$, $SD = .05$, for the proportion strategy was $M = .44$, $SD = .03$, and for the difference strategy was $M = .40$, $SD = .04$. Bonferroni post hoc test further revealed a significant difference between the base rate neglect score and both the difference ($p = .002$), and proportion score ($p = .039$). The difference of scores between the difference and proportion scores, however, was not significant ($p = .86$).

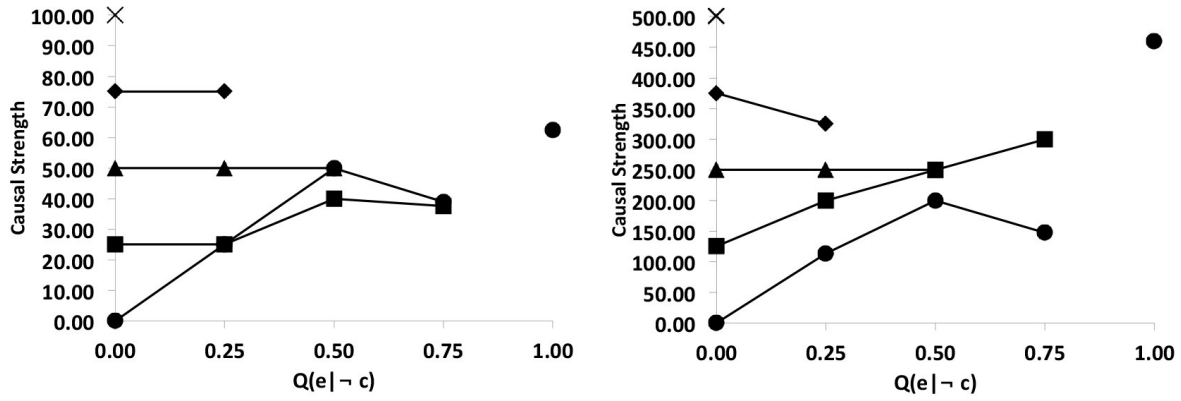


Figure 3.6: Results for Experiment 3.2. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

3.2.3 Discussion

The less consistent results between trend analysis and histogram analysis appear to be rooted in trend analysis's reliance of medians.[5] Nonetheless, in general, the majority of participants in this experiment simply neglected the base rate when making judgments. This was

clearly captured in the tendency analysis. A possible explanation for participants' neglected base rate would be because of the ambiguous link between the base rate information (i.e. control plot), and post-treatment information (i.e. experimental plot), weakening the role of the control plot: The context of the cover story was outdoors, hence allowing for other variables, such as microclimate and temperature, to influence the study and thus weakening in the link between the two plots. Participants might simply have ignored information about the control plot and focused only on the experimental plot. Further, my reassessment of the cover story suggested that if participants were focusing only on the experimental plot, they were likely to use zero as the base rate in their reasoning. This is because of one of the sentences in the cover story that says "*... an experimental plot of 100 meters square, freshly sown with corn ...*". From the word '*freshly'* in the sentence, it would make sense to imply zero as the base rate on the experimental plot. Consequently, their judgments reflected base rate neglect.

In scientific experimental design, the design of this cover story was referred to as a between-entity design, i.e. having a control and experimental entity. An alternative to this would be a within-entity design where the same subject experienced both before and after treatment situations. A parallel setting for this cover story would be to have only one seeding plot and subject it to a before-after setting. The growth on the plot could be initially measured (i.e. the base rate), and re-measured after the application of treatment. Because the same entity experienced both situations (i.e. base rate and treated), the issue of a weak link between these two magnitudes could be addressed. Thus, having a within entity design would make participants less likely to ignore the base rate information. A problem, however, was that it would be unrealistic to apply this design in this story: to let the crop grow and be measured, applying the

[5] See beginning of this chapter for an explanation of issues with trend analysis including its use of central tendency measure.

fertilizer, and making the second measurement on the already grown crops. To test this within entity design, Experiment 3.3 adopted a new cover story accommodating this design principle.

Experiment 3.3

This was the first experiment using a within-entity setting. The aim was to test whether the between-within-entity manipulation influenced which judgment strategy people would adopt, particularly with respect to the neglect of base rates during the judgment. Because it was less realistic to use the same story from Experiment 3.2 in a within-entity setting, this experiment used a new story about the influence of chemicals on algae growth.

3.3.1 Method

Participants. 88 participants recruited via Amazon Mechanical Turk, and each received a payment of $0.60 at the end of the task. Participations were limited to only those whose first language were English, and were only from the United States.[6]

Design and Procedure. This within-subject experiment utilised the same 15 conditions as the previous experiments. The cover story in this experiment began with a motivation for studying the effect of chemicals on algae growth in a natural environment. Participants were told that the study involved 15 different labs, each of which was built nearby a different natural water reservoir to study only one chemical. The lab was described to consist of an indoor pool with surface area of 100 meter square into which water from the nearby reservoir was pumped. The story continued by explaining the process of the study that began with the initial observation of algae growth inside the pool, after which the entire surface was sprayed with the chemical; two

[6] All experiments in Chapter 3 and 4 that employed Amazon Mechanical Turk went online under different Batch (hence different time) but using the same Project (which could also be referred to job/task). Because these experiments were in the same Task, only unique participants were able to participate across experiments despite different batch (i.e. time). This was the control mechanism to avoid duplication of participation across experiments.

weeks after this, a second observation of algae growth was then taken (see Appendix A for the cover story).

Participants proceeded with the first condition after going through the story. The information for each condition, i.e. the area covered by algae before and after treatment, was presented in the same visual format as in the previous experiment: Three columns where the right column portrayed the area before treatment with the chemical, while the left column portrayed the area after the treatment. The middle column was reserved for the legend showing the surface area if it was completely filled with algae, and if it was completely empty. The question underneath was "*Now imagine another pool filled with water from a different lake and a surface area of 100 meters square. None (i.e. 0 square meters) of the 100 square meter surface of the pool is covered with algae. If we apply this chemical substance on this new pool, how much of its area would be covered by algae after two weeks?*" Participants answered this by typing in an empty box that only allows numbers from 0 to 100. Following this question was the second-limit question worded exactly the same but replaced with "*500 meters square*" whenever there was "*100 meter square*". After these questions, participants proceed with the next condition, in the same format.

3.3.2 Results

Figure 3.2 shows a skewed distribution of the data. Thus, the trend analysis used medians in the plots (Figure 3.7). Comparing this plot with prediction plots (Figure 3.1), no single strategy was evident to be the dominant. In the consistent limit plot, only conditions $C_{0.75}$ and $C_{0.50}$ could be considered flat, suggesting use of the difference strategy. While judgments of conditions $C_{0.00}$ followed strictly a base rate neglect pattern, judgments of conditions $C_{0.25}$ had indication of both difference and base rate neglect. For the higher limit, the pattern was cleaner,

where both conditions $C_{0.25}$ and $C_{0.50}$ were obviously reflecting difference strategy. Conditions $C_{0.75}$ and $C_{0.00}$ in contrast, did not fit with any prediction. Further analysis using Friedman's Test on the judgments revealed significant results for all groups in the consistent limit case ($C_{0.00}$: $\chi^2(3) = 109.844$, $p \leq .001$; $C_{0.25}$: $\chi^2(3) = 77.411$, $p \leq .001$; $C_{0.50}$: $\chi^2(2) = 47.294$, $p \leq .001$; $C_{0.75}$: $\chi^2(1) = 8.018$, $p = .005$), as well as in the higher limit ($C_{0.00}$: $\chi^2(3) = 83.489$, $p \leq .001$; $C_{0.25}$: $\chi^2(3) = 76.365$, $p \leq .001$; $C_{0.50}$: $\chi^2(2) = 22.691$, $p \leq .001$; $C_{0.75}$: $\chi^2(1) = 6.582$, $p = .010$).

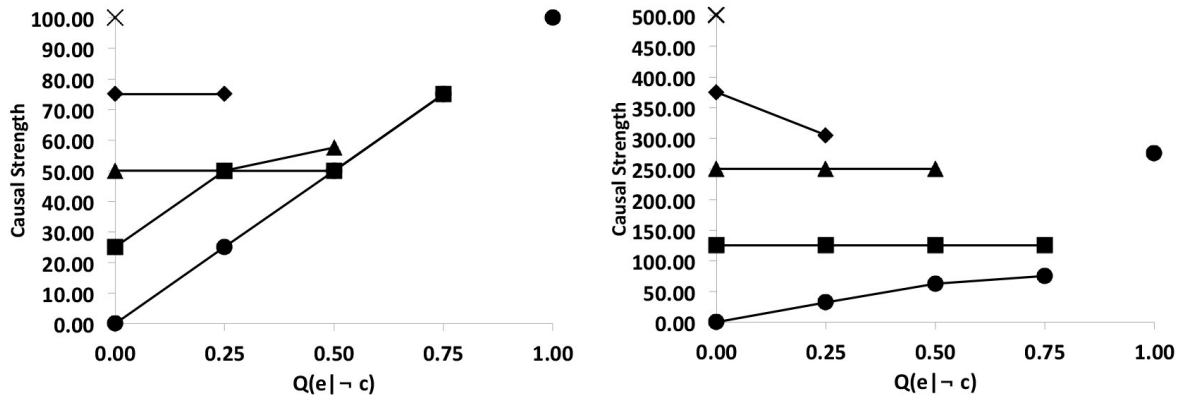


Figure 3.7: Results for Experiment 3.3. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

The histogram analysis revealed a complex pattern of results: Excluding five fully overlapping conditions, six of the conditions were dominated by base rate neglect strategy, and four of the conditions were dominated by difference strategy. The proportion strategy dominated in two conditions that were partially overlapped with base rate neglect. Tendency analysis in the data revealed a significant main effect of the three scores, $F(1.12, 97.00) = 5.49$, $p = .018$ after Greenhouse-Geisser correction. Base rate neglect had the highest mean score ($M = .53$, $SD = .03$), whereas proportion strategy had the lowest ($M = .42$, $SD = .02$). Difference strategy meanwhile had mean score of $M = .46$, $SD = .03$. Further Bonferroni *post hoc* tests, however, unveiled that only the difference of scores between base rate neglect and proportion was

significant ($p \leq .001$), whereas the others were not ($p = .340$ for difference of base rate neglect score and difference score, and $p = .465$ for difference of proportion score and difference score).

3.3.3 Discussion

Similar to Experiment 3.1, the significant results of Friedman's Test to the flat lines of conditions $C_{0.75}$ and $C_{0.50}$ in the consistent limit and conditions $C_{0.25}$ and $C_{0.50}$ in the higher limit were perplexing. Further inspection of the distributions (c.f. Figure 3.2) revealed that they were bi-modal. Because the test relies on central tendency (i.e. mean of the ranks of the judgments), it thus poorly captured the results.

Base rate neglect, while still the dominant strategy in this experiment, was less intense than in Experiment 3.2. This suggested that a within-entity setting provided better link between base rate and treated magnitudes because both magnitudes information came from the same entity. Another competing explanation for the results would be the influence of context, which was algae growth in this experiment instead of crops growth in Experiment 3.2. In this experiment (algae context), the situation was described as taking place inside laboratories instead of outdoors as in Experiment 3.2's crops growth context, hence other variables were better controlled. To address this competing argument, Experiment 3.4 adopted this same algae context but using a between-entity setting as in Experiment 3.2. To deal with variability that could be associated with other uncontrollable factors, such as microclimate and temperature in Experiment 3.2, the cover story described that both control and experimental pools were using water from the same source.

Experiment 3.4

This experiment was a replication of Experiment 3.3, but using a between-entity setting. The objective of this experiment was to study whether situation (within-entity vs. between-entity) or context was more influential to the judgments, in particular the base rate neglect.

3.4.1 Method

Participants. 18 undergraduate students from the School of Psychology, Cardiff University participated to fulfil part of a course requirement. The same exclusion criteria from Experiment 3.1 were adopted.

Design and Procedure. This experiment replicated Experiment 3.3 except with some changes on the cover story to reflect the between-entity situation. Specifically, the story informed participants that within each lab, there were two indoor pools (instead of only one pool in Experiment 3.3): one pool was reserved as control and received no treatment, whereas the other pool was sprayed with the chemical. After going through the story, participants proceeded with the conditions that were presented in the same format as in Experiment 3.3. In each condition, participants also had to answer two questions of which the first question corresponded to the consistent limit with the story (100 meter square), while the second question used a higher limit (500 meter square).

3.4.2 Results

Trend analysis used medians in the plots because the distribution was skewed (see Figure 3.8). For the consistent limit, the plot clearly indicated base rate neglect in all conditions. For the higher limit, the trend was more complex: besides a trend for conditions $C_{0.75}$, which hinted at base rate neglect, and conditions $C_{1.00}$, which hinted at a strict difference strategy, the remaining trends were not clear. Friedman's Test on the judgments for both limits suggested agreement with

the observation as all results were significant: for the consistent limit, $C_{0.00}$: $\chi^2(3) = 16.339$, $p = .001$; $C_{0.25}$: $\chi^2(3) = 18.496$, $p \leq .001$; $C_{0.50}$: $\chi^2(2) = 11.261$, $p = .004$; and $C_{0.75}$: $\chi^2(1) = 12.000$, $p = .001$; and for the higher limit, $C_{0.00}$: $\chi^2(3) = 20.042$, $p \leq .001$; $C_{0.25}$: $\chi^2(3) = 11.396$, $p = .010$; $C_{0.50}$: $\chi^2(2) = 7.091$, $p = .029$; and $C_{0.75}$:$\chi^2(1) = 11.000$, $p = .001$.

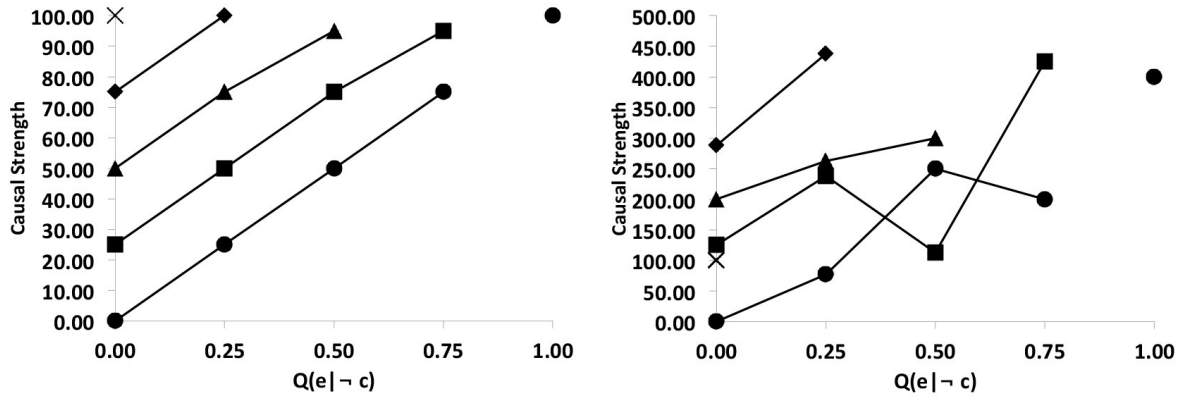


Figure 3.8: Results for Experiment 3.4. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

Histogram analysis produced results that also supported base rate neglect as the dominant strategy in all conditions. As for the higher limit, the base rate neglect judgments were observed in two ways: either in a strict neglect (e.g. neglected 25 m^2, and opted 75 m^2 in [0.25:0.75] condition), or scaled neglect (e.g. neglected 125 m^2, and opted 375 m^2 in [0.25:0.75] condition). Tendency analysis on the data indicated a non-significant main effect of the scores after Greenhouse-Geisser correction, $F(1.08, 16.15) = 2.19$, $p = .157$.

3.4.3 Discussion

Even though trend and histogram analyses suggested that base rate neglect was the dominant strategy, tendency analysis showed that this main effect was non-significant. Thus, in this experiment, all strategies were equally competing to be the most prominent. Even though the same algae context was used in both this experiment and Experiment 3.3, the inconsistency of

the results across these experiments indicated the influence of having the information in between-entity, instead of within-entity as in Experiment 3.3, over the context. On the other hand, despite adopting the same between-entity situation like in Experiment 3.2 the inconsistent result of this experiment with that experiment suggested the influence of context, over situation (i.e. algae vs. crop). Nonetheless, the results of this experiment still showed relatively high score of base rate neglect.

Another way of looking at base rate neglect was that participants were simply copying the magnitude of the effect after receiving treatments. This act of copying was the simplest strategy because of the base rate used in the question was the same as in the cover story. Including this experiment, all experiments used this setting. Thus, to go forward with the investigation, I continued the following experiment by making the base rates in the cover story and in the questions dissimilar.

Experiment 3.5

This experiment adopted the same design as in Experiment 3.3, i.e. within-entity. The use of within- instead of between-entity was because in within-entity situation, the story was simpler and less cognitive demanding. By making the base rate in the cover story and in the question to be dissimilar, participants were required to put more effort when making judgment. This probably would address the issue of base rate neglect that was noteworthy in previous experiments.

3.5.1 Method

Participants. 63 participants, recruited via Amazon Mechanical Turk, participated for a small reimbursement ($0.60). The same exclusion criteria from Experiment 3.3 were adopted in this experiment.

Design and Procedure. This experiment duplicated the same Experiment 3.3 with a new manipulation in the judgment question. The new question was "*Imagine a pool filled with water from a different lake and a surface area of 100 meters square. Further, imagine that when you arrive at the site, 25 square meters of the 100 square meter surface of the pool is already covered with algae. If we apply this chemical substance on this new pool, how much of its area would be covered by algae after two weeks?*" In this new question, participants were informed that the base-rate of the counterfactual pool was 25 meter square, instead of 0 meter square in all previous experiments. The rest of the procedure was exactly the same with Experiment 3.3.

3.5.2 Results

Skewed distribution of the data required the use of medians in the trend analysis (Figure 3.9). For the consistent limit, qualitative observation of the plot suggested that participants used both difference and proportion strategies. In particular, the difference strategy dominated conditions $C_{0.00}$ and $C_{0.25}$, whereas the proportion strategy dominated conditions $C_{0.75}$. The trend for conditions $C_{0.50}$, however, did not match any of the predictions. In the higher limit conditions, the strategy suggested by the trends was less obvious. While trends for conditions $C_{0.00}$ and $C_{0.75}$ followed their consistent limit counterparts, trends for the remaining conditions were less obvious.

Applying Friedman's Test on the consistent limit judgments revealed that results for conditions $C_{0.25}$ ($\chi^2(3) = 18.337$, $p \leq .001$)[7] and $C_{0.75}$ ($\chi^2(1) = 3.846$, $p = .050$) were significant, whereas results for conditions $C_{0.00}$ ($\chi^2(3) = 7.146$, $p = .067$) and $C_{0.50}$ ($\chi^2(2) = .970$, $p = .616$) were not. This pattern was also applicable in the higher limit where significant results were for

[7] Inspecting the distribution of these conditions revealed that modal judgments correspond to 50 meter square, but the minority judgments exhibit outcome density bias, which is picked up by Friedman's Test to produce a significant result even though the trend was clearly a flat line in Figure 3.5. For further details on this bias refer to Discussion of Experiment 3.1

conditions $C_{0.25}$ ($\chi^2(3) = 8.380$, $p = .039$) and $C_{0.75}$ ($\chi^2(1) = 3.908$, $p = .048$), and non-significant results were for conditions $C_{0.00}$ ($\chi^2(3) = 3.843$, $p = .279$) and $C_{0.50}$ ($\chi^2(2) = 1.943$, $p = .379$).

Histogram analysis was mainly on condition [0.50:1.00] because the mode in this histogram did not fit with any strategies under current study. Further inspection of the results suggested that participants adopted a new reasoning strategy – multiplication. This strategy is discussed in more detail below. Re-conducting histogram analysis to consider multiplication strategy revealed that difference seemed to be the dominant strategy, but closely followed with multiplication strategy. Except in condition [0.50:1.00], judgments in other conditions matched with difference strategy predictions.

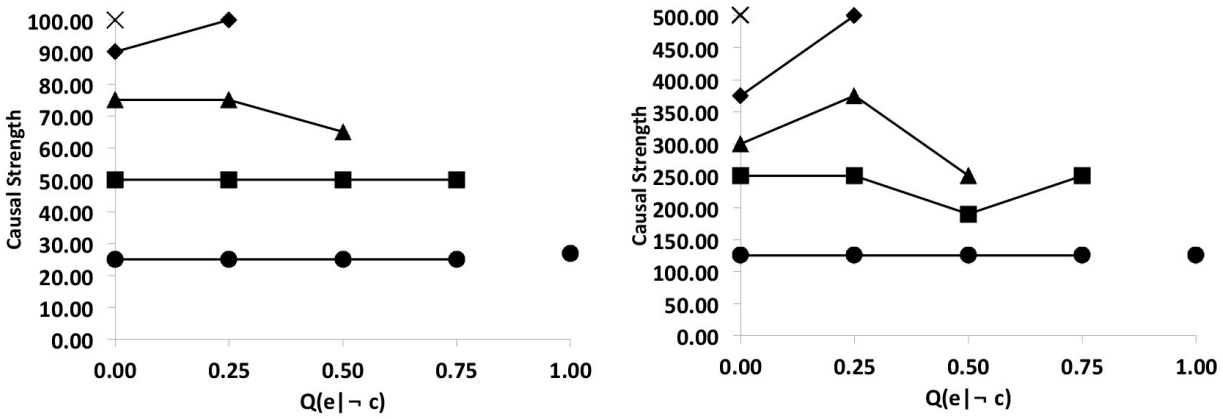


Figure 3.9: Results for Experiment 3.5. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

Tendency analysis on this data included multiplication strategy as well. Thus, instead of considering scores from only three strategies (i.e. base rate neglect, proportion, and difference), this analysis considered four strategies. Greenhouse-Geisser corrected test found a significant main effect of these scores, $F(1.73, 107.07) = 17.04$, $p \leq .001$. The mean score for base rate neglect was $M = .37$, $SD = .03$, for proportion strategy was $M = .39$, $SD = .03$, for difference strategy was $M = .54$, $SD = .04$, and for multiplication strategy was $M = .50$, $SD = .04$. Further

Bonferroni post hoc test revealed that difference of mean scores between base rate neglect and proportion strategies was not significant ($p \geq .05$), unlike the rest: base rate neglect vs. difference ($p \leq .001$), base rate neglect vs. multiplication ($p = .003$), proportion vs. difference ($p \leq .001$), proportion vs. multiplication ($p \leq .001$), and difference vs. multiplication ($p = .024$).

3.5.3 Discussion

The new strategy was found during scrutiny of condition [0.50:1.00]. In this condition, neither proportion's nor difference's predictions could best describe the mode of the distribution. When given an increase of the growth area from 50 to 100 m^2 in the presence of the candidate cause, participants might have reasoned that the candidate cause would double the area. Therefore, when given 25 m^2 initially in the counterfactual pool, the majority of participants reasoned that the same candidate cause should also double the area from 25 to 50 m^2. I referred to this new reasoning approach as multiplication strategy, which as far I am aware, has no precedence in the causal learning literature.

One noteworthy observation in this experiment was the massive reduction of base rate neglect relative to previous experiments. A possible explanation for this was that those base rate neglect judgments were substituted with judgments using multiplication strategy. In previous experiments, when participants were asked to judge in counterfactual situations with zero base rate, they could not. This is because, any multiplication-based judgments with zero base rates would result in zero growth, indicating incapacity of the candidate cause to produce the effect, which contradicts with the observed evidence that the candidate cause, to some degree, does generate the effect magnitudes. Thus, for multiplication strategy to be relevant, the base rate provided in the counterfactual scenario could not be zero. In other words, the absence of the opportunity to reason using a multiplication strategy in previous experiments had triggered

participants to adopt other strategies instead, and perhaps base rate neglect was the easiest to do. In contrast, when they had opportunity to channel multiplication-based judgments in the non-zero base rate as in this experiment, base rate neglect judgments massively went down. More detail on multiplication strategy is in the following section. Even though the multiplication strategy was plausible in this experiment, in general, the difference strategy was the most dominant across the conditions.

3.5a Multiplication Strategy

Following the same example from Chapter 1 and the beginning of this chapter – i.e. the example of algae covering the surface area of a new pool after receiving treatment with mineral C – another plausible prediction besides 60 m^2 (difference strategy) and 70 m^2 (proportion strategy) is 30 m^2: Reasoners might consider the efficacy of mineral C in terms of its capacity to multiply the covered surface area relative to before treatment. A reasoner who adopts this strategy learned that in scenario 3C mineral C tripled the area of algae growth from 25 m^2 to 75 m^2. Therefore in scenario 5, with the same efficacy, mineral C should also triple the algae growth from 10 m^2 to 30 m^2.

Multiplication strategy stems from an interaction between background causes and candidate cause in producing the effect instead of considering the candidate cause as directly changing the effect. In other words, multiplication based reasoners might have conceived of algae growing on their own (due to the background causes), with the candidate cause merely amplifying this tendency, rather than acting as a cause on its own. In this case, the candidate cause influences the propensity of the background causes to produce the effect. This violates the normative assumptions of a proportional framework, namely that the influence of background

causes on the effect should remain constant before and after treatment, and that the cause and background each influence the effect independently.

One important issue with this strategy was involving zero base rates. As an example, in the absence of the candidate cause, algae coverage of the pool was 0 m^2 (i.e. no algae were present), and increased to 20 m^2 when the candidate cause was administered. Multiplicative reasoners cannot compute the efficacy index of the candidate because its interaction with background causes is less clear when the base rate is zero. Besides this issue of zero base rates, an immediate question pertaining to this strategy was that whether it is applicable only in the algae-mineral context. Experiment 3.6 aimed to answer this question.

Experiment 3.6

Given the newly found multiplication strategy in Experiment 3.5, the investigation progressed by examining whether this strategy was applicable in other context as well. In Experiment 3.5, the context was about algae growth. Because algae growth was a natural process, the background causes would continue to keep on producing algae growth. In the presence of the candidate cause, it was also natural to think that the candidate cause acts on the occurring background cause (i.e. a causal interaction) to produce algae growth. Because the multiplication strategy exemplifies an interaction between background and candidate causes, hence it was prevalent in the algae story. To test whether the multiplication strategy is valid only in these interaction-based contexts, Experiment 3.6 replicated Experiment 3.5 but using a different story in which the influence of background causes on effect was more consistent throughout.

3.6.1 Method

Participants. 63 participants, recruited via Amazon Mechanical Turk, participated for a small amount of reimbursement ($0.60). The exclusion criteria from Experiment 3.3 were also implemented for this experiment.

Design and Procedure. This experiment also employed the same conditions as the previous experiments. The cover story introduced the influence of 15 chemical additives (corresponded to the 15 conditions) on the runniness of engine oils, each studied separately in different labs. The story described the study to follow a drip-test procedure in which 5 grams of oil were deposited on one end of a 10 cm long test slate, slanted at an angle of 45 degrees. After an interval of 5 minutes the total length (out of 10 cm) travelled by the drop of oil was measured as an indicator of runniness: The greater the distance the oil travelled, the greater the runniness of oil. The story proceeded by mentioning that for each additive, the scientists always tested the runniness of oil before adding the additives, and then mixed an additive into the oil and repeated the test (see Appendix A for the cover story).

After going through the story, participants received the information from a condition about the length of oil travelled down the test slate. The left side portrayed the travel distance of oil without any additives, whereas the right side displayed the distance of oil with additive under investigation. Underneath these information was question "*Imagine that we performed a test with the same engine oil at a different location. As before, we drip 5 gram of the oil on the raised end of a 10 cm long test slate, angled at 45 degrees. After 5 minutes, the oil has traveled 2 cm down this slate. If we would mix this additive into the oil and repeated the test under the same conditions, how far down the slate would the oil travel now (after 5 minutes)?*" In this

experiment, however, only one question was used. After this question, participants proceeded with the next condition.

3.6.2 Results

Trend analysis used medians in the plot because results' distributions were skewed (Figure 3.10). Qualitative observation of the plot suggested a mixture of difference and multiplication as the most influential strategies. Difference strategy was particularly evident in conditions along the vertical axis, as well as in condition [0.75:0.50] and [1.00:0.75]. Friedman's Test showed that all conditions (i.e. conditions $C_{0.00}$: $\chi^2(3) = 25.769$, $p \leq .001$;[8] conditions $C_{0.25}$: $\chi^2(3) = 8.994$, $p = .029$; and conditions $C_{0.50}$: $\chi^2(2) = 19.037$, $p \leq .001$) were significant except for conditions $C_{0.75}$ ($\chi^2(1) = .333$, $p = .564$).

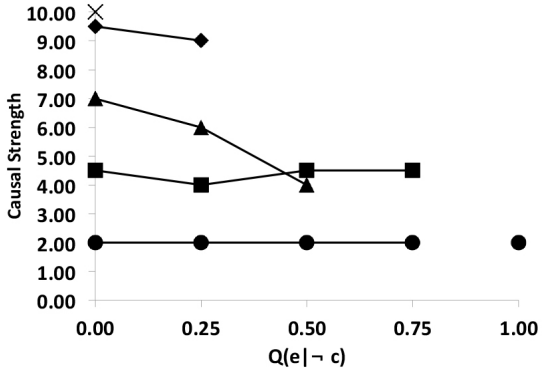


Figure 3.10: Results for Experiment 3.6.

Results from histogram analysis also supported these results. In conditions where multiplication strategy was possible, (i.e. non-zero base rate conditions including [0.75:0.50] and [1.00:0.75]), multiplication strategy dominated all conditions except in condition [0.75:1.00], where difference strategy took over. Tendency analysis on the data, corrected using Greenhouse-

[8] Even though the trend in Figure 3.6 for these conditions was clearly a flat line, inspection of the histogram revealed that these significant results were due to minority judgments exhibiting outcome density bias. For further details, refer to Discussion of Experiment 3.1

Geisser, revealed a significant main effect of the four scores $F(1.38, 87.09) = 71.58$, $p \leq .001$. The mean score for base rate neglect was $M = .13$, $SD = .01$, for proportion strategy was $M = .38$, $SD = .02$, for difference strategy was $M = .38$, $SD = .02$, and for multiplication strategy was $M = .59$, $SD = .04$. Bonferroni *post hoc* test further unveiled that all pairwise comparison were significant except between proportion and difference scores ($p \geq .05$): base rate neglect vs. proportion ($p \leq .001$), base rate neglect vs. difference ($p \leq .001$), base rate neglect vs. multiplication ($p \leq .001$), proportion vs. multiplication ($p \leq .001$), and difference vs. multiplication ($p \leq .001$).

3.6.3 Discussion

In general, multiplication strategy dominated this experiment except in conditions on the vertical axis. In these conditions, the predicted values for multiplication strategy were infinity, which was not captured in this experiment. Thus, in these conditions, participants opted for judgments that fit both the difference and proportion predictions.

This experiment considered the possibility of context dependency of algae-mineral. This is because, the setting of algae-mineral context involving the nature might be more accommodating to thinking of an interaction between candidate cause and background causes, of which the core of multiplication based reasoning. By adopting the lab based setting as engine oil-additive context in this experiment, the influence of background causes on the effect was more controlled. Thus participants might perceive the influence of background causes on the effect to be more stable throughout the process of presenting the effect in the presence and absence of candidate cause. Therefore, in this context, I predicted that multiplication strategy would be less appealing to the participants.

The results of this experiment, however, indicated that this prediction did not hold: evidence for multiplication-based reasoning was also prominent in this experiment. Interestingly, with the increase of support towards multiplication strategy, judgments neglecting the base rates were plummeting. While there was no direct evidence to associate this decrease with multiplication strategy, the inverse relationship between multiplication-strategy and base rate neglect was also observable in Experiment 3.5. Nonetheless, this signals the relevance of the multiplication strategy in reasoning about causal relation involving continuous outcomes.

3.6a Generative vs. Preventive

The previous six experiments were dealing with causal relations in generative scenarios. The preventive version of these experiments (except Experiment 3.2 and 5) aimed to search for the most prominent reasoning strategy in a preventive scenario. In this scenario, the observable effect magnitude in the presence of candidate cause was always smaller, or less, than its magnitude in the absence of the cause (i.e. the base rate). Thus, the cause should be perceived to lower the effect magnitude.

3.6a.1 Strategies Recap in Preventive Scenario

In preventive scenarios, all strategies are also rooted in the same idea as in generative scenario, but with some minor differences. For the difference strategy, computing the strength index was the same as in generative. However, because the magnitude of the effect when the candidate cause was present was always smaller than when the cause was absent, the strength index of the difference strategy carries a negative sign in front of the value. As for the proportion strategy, the idea was to consider the proportion of cause efficacy relative to its potential maximum efficacy. In the preventive scenario, the maximum efficacy of a candidate cause was to completely prevent the effect. Therefore, the upper limit of effect magnitude is not a concern in

preventive scenarios when considering proportion strategy. In other words, a candidate cause that has a maximum preventive power would be able to completely wipe out any effect regardless of the original value (i.e. base rates). Multiplication-based reasoners would consider an interaction between candidate cause and background causes when preventing the effect. Similar to the generative scenario, for the preventive scenario, the multiplication index was computed by considering the ratio between effect magnitudes when the cause is present vs. when it is absent. For complete preventive predictions of the 15 conditions, see Table 1.2. It is clear in the table that the predictions for multiplication and proportion strategy always overlap. The reasons for this as well as implications are discussed in Appendix C.

3.6a.2 Preventive Experiments Overview

Four experiments in preventive scenarios mirrored the generative counterpart: Experiment 3.7 with 3.1, Experiment 3.8 with 3.3, Experiment 3.9 with 3.4, and Experiment 3.10 with 3.6.

Experiment 3.7

This experiment is a parallel to Experiment 3.1. The aim was to study the prominent reasoning strategy of causal relations with continuous outcomes in preventive scenario.

3.7.1 Method

Participants. 30 undergraduate students from the School of Psychology, Cardiff University participated to fulfil part of a course requirement. I adopted the same exclusion criteria from Experiment 3.1 in this experiment.

Design and Procedure. The experiment adopted the same 15 conditions but with the role of the quantity as reversed (see Table 1.2) – effect quantity when the cause was absent in generative became quantity when the cause was present in this experiment. Participants in this

experiment went through the same skin rash context with Experiment 3.1 cover story except with some changes to reflect the preventive scenario. In particular, the story focuses on the influence of ointment in preventing skin rash, instead of on the side effect of cosmetic cream in generating skin rash in Experiment 3.1. The other settings of the story were consistent – 15 different labs, each investigating one ointment on a 10 centimetres square skin rash. Participants then received the condition information in the same visual format as in Experiment 3.1, followed by the question "*Now imagine a new allergy patient suffering from a rash of 10 centimeters square. If we apply the ointment, how large would the area of rash be ?*". The second question used the same wording except with change on the skin area to be *50 centimeters square*. After making judgments in both question for that condition, participants moved on to the next condition.

3.7.2 Results

The data were skewed, requiring trend analysis to use medians (Figure 3.11). For the consistent limit, observing the plot suggested proportion/multiplication as the dominant strategy. This trend continued in the higher limit as well. Further, judgments for conditions with maximum effect magnitude (i.e. conditions [1.00:1.00], [1.00:0.75], [1.00:0.50], and [1.00:0.25]) did not reach the maximum upper boundary as predicted by proportion/multiplication. Friedman's Test further supported the observation that proportion/multiplication was dominant with the significant results for all conditions (i.e. conditions $C_{0.25}$: ($\chi^2(3) = 57.854$, $p \leq .001$), conditions $C_{0.50}$: ($\chi^2(2) = 15.892$, $p \leq .001$), and conditions $C_{0.75}$: ($\chi^2(1) = 9.738$, $p = .002$) except conditions $C_{0.00}$ ($\chi^2(3) = 4.500$, $p = .212$). This is also evident in the higher limit (conditions $C_{0.00}$: ($\chi^2(3) = 1.222$, $p = .748$), conditions $C_{0.25}$: ($\chi^2(3) = 27.931$, $p \leq .001$), conditions $C_{0.50}$: ($\chi^2(2) = 12.302$, $p = .002$), and conditions $C_{0.75}$: ($\chi^2(1) = 3.846$, $p = .050$)).

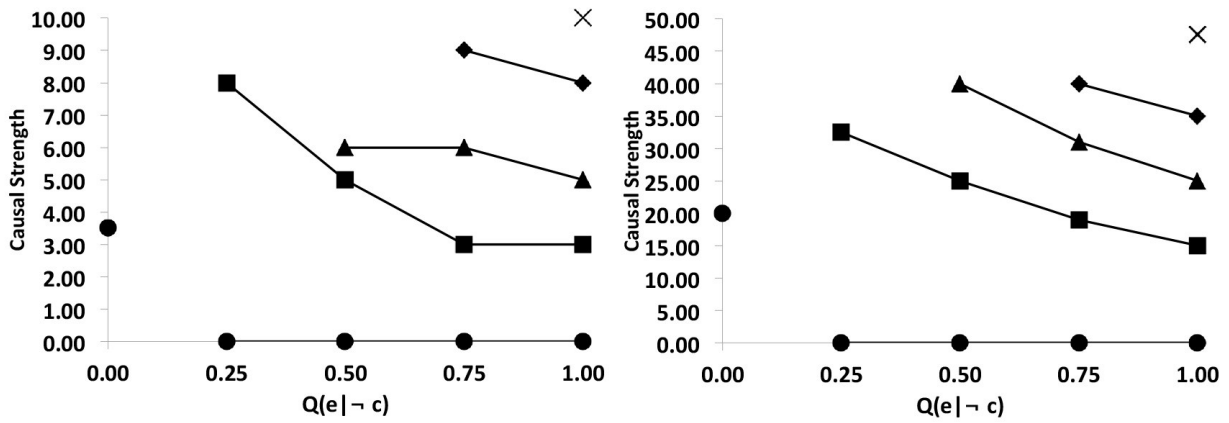


Figure 3.11: Results for Experiment 3.7. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

Results from histogram analysis also supported that proportion/multiplication was the dominant strategy. In all conditions, the modes reflected proportion/multiplication strategy. Tendency analysis on the data revealed a significant main effect of the scores, $F(2,58) = 34.85$, $p \leq .001$. The mean scores for base rate neglect was $M = .31$, $SD = .02$, for proportion/multiplication strategy $M = .52$, $SD = .04$, and for difference strategy $M = .41$, $SD = .03$. Bonferroni *post hoc* test further revealed that all pairwise differences of these scores were significant: base rate neglect vs. proportion/multiplication ($p \leq .001$), base rate neglect vs. difference ($p = .003$), and proportion vs. difference ($p \leq .001$).

3.7.3 Discussion

All analyses pointed to proportion/multiplication as the most prominent strategy in this experiment. Even though this experiment used the same cover story as in Experiment 3.1, the results were different. This inconsistency was perhaps attributable to the different structure between generative and preventive (for further discussion on this structural difference refer to the General Discussion).

Unlike in generative scenario, the concern regarding the plausibility of the cover story was not the issue in preventive scenario. This is because the candidate cause in the story was about ointments that had effectiveness to reduce the magnitude of skin rash. Therefore, when the story mentioned that there was already skin rash at the beginning of the study, applying the ointments on top of it was completely sensible.

Experiment 3.8

This experiment continued the investigation in the preventive scenario using the algae growth context. Parallel to Experiment 3.3, the cover story also involved a within-entity situation.

3.8.1 Method

Participants. 89 participants, recruited via Amazon Mechanical Turk, participated for a small reimbursement ($0.60). This experiment employed the same exclusion criteria from Experiment 3.3.

Design and Procedure. Participants experienced the same 15 conditions as in previous experiments. They also went through the same cover story as in Experiment 3.3 but with a very minor modification – explicitly a substitution of any word 'cause' with word 'prevent'. The presentation of the conditions was also using the same visual template. Underneath the information, participants received the question: "*Now imagine another pool filled with water from a different lake and a surface area of 10 m^2. The entire 10 m^2 surface of the pool is already covered with algae. If we apply this chemical substance on this new pool, how much of its area would be covered by algae after two weeks?*" The second question was exactly the same except with 50 m^2 surface area. The next condition appeared on the screen after participants made judgments for that condition.

3.8.2 Results

The trend analysis was using medians because of the skewed distribution of the data (Figure 3.12). For the consistent limit, the trend clearly matched proportion/multiplication predictions. Similarly in the higher limit, the trend also reflected proportion/multiplication strategy except with some outcome density effect in the conditions $C_{0.00}$.[9] Because of this, Friedman's Test on all conditions revealed significant results in the consistent limit (conditions $C_{0.00}$: ($\chi^2(3) = 60.019$, $p \leq .001$), conditions $C_{0.25}$: ($\chi^2(3) = 162.018$, $p \leq .001$), conditions $C_{0.50}$: ($\chi^2(2) = 107.553$, $p \leq .001$), and conditions $C_{0.75}$: ($\chi^2(1) = 58.778$, $p \leq .001$)); and similarly in the higher limit (conditions $C_{0.00}$: ($\chi^2(3) = 52.083$, $p \leq .001$), conditions $C_{0.25}$: ($\chi^2(3) = 165.295$, $p \leq .001$), conditions $C_{0.50}$: ($\chi^2(2) = 93.483$, $p \leq .001$), and conditions $C_{0.75}$: ($\chi^2(1) = 45.762$, $p \leq .001$)).

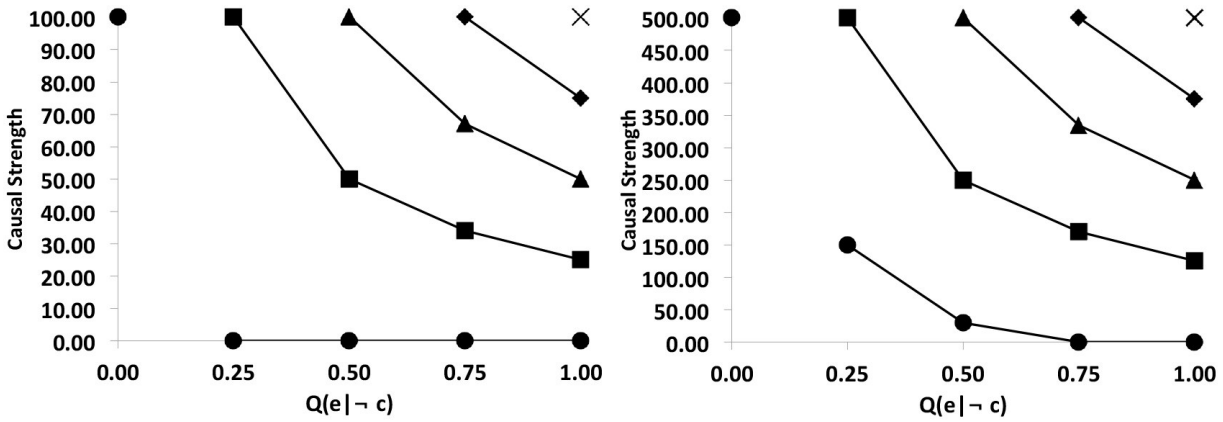


Figure 3.12: Results for Experiment 3.8. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

Histogram analysis revealed results that also suggested that proportion/multiplication as the dominant strategy. Modes in all conditions represented this strategy. Tendency analysis on the data, corrected with Greenhouse-Geisser, showed a significant main effect of the scores, $F(1.46,$

128.28) = , $p \leq .001$. On average, base rate neglect scored $M = .57$, $SD = .03$, while proportion/multiplication strategy scored $M = .58$, $SD = .03$; the difference strategy had the lowest mean score at $M = .42$, $SD = .02$. Further Bonferroni *post hoc* test unveiled that these mean scores of base rate neglect and proportion/multiplication strategy were not significantly different ($p \geq .05$). Meanwhile, the differences of mean scores of other pairs were significant: base rate neglect and difference strategy ($p \leq .001$), and proportion and difference strategy ($p \leq .001$).

3.8.3 Discussion

The results of the tendency analysis were slightly inconsistent with the other two analyses where the tendency score of proportion/multiplication was not significantly different with the score of base rate neglect. Nonetheless, considering the results of all of the analyses, proportion/multiplication dominated over the other strategies. Another noteworthy observation was about the base rate neglect judgments. Despite having both cover stories about a within-entity design, the mean scores of base rate neglect were higher in this experiment than in Experiment 3.7. This difference may be attributed to the different cover story between these two experiments: cream-rash in Experiment 3.7, and chemical-algae in this.

Experiment 3.9

The investigation continued with the preventive counterpart of Experiment 3.4. The objective was to study whether a between-entity situation has any influence on the reasoning strategy.

---

[9] Refer to Discussion of Experiment 3.1 for details.

3.9.1 Method

Participants. 17 undergraduate students from the School of Psychology, Cardiff University participated to fulfil part of a course requirement. Participations for this experiment was subjected to the same exclusion criteria as in Experiment 3.1.

Design and Procedure. This experiment replicated Experiment 3.8 but with modifications to show a between-entity situation. Specifically, the story informed participants that within each lab, there were two indoor pools (instead of only one pool in Experiment 3.8): one pool was reserved as control and received no treatment, whereas the other pool was sprayed with the chemical. Participants also received the condition information in the same visual format and content as in Experiment 3.8.

3.9.2 Results

Skewed distribution in the data required that trend analysis used medians in the plots (Figure 3.13). In both limits, qualitative observation suggested that the trends reflected base rate neglect. This is supported by Friedman's Test that produced significant results for all conditions in the consistent limit (conditions $C_{0.00}$: $\chi^2(3) = 29.956$, $p \leq .001$; conditions $C_{0.25}$: $\chi^2(3) = 41.759$, $p \leq .001$; conditions $C_{0.50}$: $\chi^2(2) = 21.344$, $p \leq .001$; and conditions $C_{0.75}$: $\chi^2(1) = 9.941$, $p = .002$), as well as in the higher limit (conditions $C_{0.00}$: $\chi^2(3) = 17.890$, $p \leq .001$; conditions $C_{0.25}$: $\chi^2(3) = 43.148$, $p \leq .001$; conditions $C_{0.50}$: $\chi^2(2) = 15.180$, $p \leq .001$; and conditions $C_{0.75}$: $\chi^2(1) = 6.250$, $p = .012$).

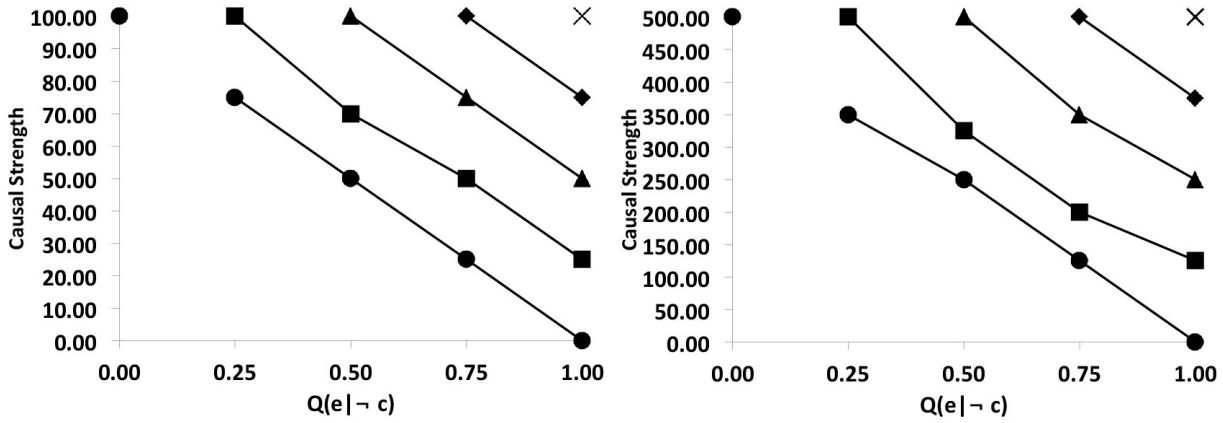


Figure 3.13: Results for Experiment 3.9. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

Histogram analysis also produced the same results that participants mostly neglected the base rate. Tendency analysis on the data also showed a significant main effect of strategy after Greenhouse-Geisser correction, $F(1.18, 18.82) =$, $p \leq .001$. The highest mean score was for base rate neglect at . $M = .67$, $SD = .08$, followed by proportion/multiplication strategy at . $M = .50$, $SD = .06$. Meanwhile, the difference strategy had the lowest mean score of $M = .33$, $SD = .04$. Bonferroni *post hoc* test further revealed that all pairwise differences between these scores were significant: base rate neglect vs. proportion ($p = .014$), base rate neglect vs. difference ($p = .001$), and proportion vs. difference ($p \leq .001$).

3.9.3 Discussion

All analyses showed that most judgments followed a base rate neglect trend. Even though the context of this experiment was the same as in Experiment 3.8, having the information presented in a between-entity situation increases participants' tendency for neglecting the base rates. This result was similar to the generative counterpart (Experiment 3.4). As discussed in Experiment 3.2 and 3.3, having a within-entity situation strengthens the link between base rate and treated magnitudes because the information corresponds to the same entity. In contrast, the

link between the information was weaker in between-entity situations because the setting allows more room for assumptions about other uncontrollable factors to influence the magnitudes. If these base rate judgments were discarded, the next strategy that participants adopted was proportion/multiplication. The tendency analysis suggested that this strategy was significantly more dominant than the difference strategy of which was the least adopted strategy in preventive scenario.

Experiment 3.10

This experiment was the preventive parallel of Experiment 3.6, aimed to study whether, in a different context, proportion/multiplication was still the most prominent strategy, and whether difference was the least relevant strategy. In this study was a new story about additives that influence liquidity of engine oils.

3.10.1 Method

Participants. 61 participants recruited from Amazon Mechanical Turk participated for a small payment of $0.60. Participants in this experiment went through the same exclusion criteria as in Experiment 3.3.

Design and Procedure. This experiment was a replication of Experiment 3.6 with changes to transform it into a preventive scenario. There was no change to the cover story in this experiment, as it was generic enough for both generative and preventive scenario. The presentation of condition information also used the same format, except the question was modified to be "*Imagine that we performed a test with the same engine oil at a different location. As before, we drip 5 gram of the oil on the raised end of a 10 cm long test slate, angled at 45 degrees. After 5 minutes, the oil has travelled 10 cm down this slate. If we would mix this additive into the oil and repeated the test under the same conditions, how far down the slate*

would the oil travel now (after 5 minutes)?" Similar to Experiment 3.6, this experiment asked participant only one question.

3.10.2 Results

The distribution of the data was skewed. Thus, medians were used in the trend analysis (Figure 3.14). Qualitative observation of the plot indicated that proportion/multiplication was the dominant strategy. In all condition groups, the trend followed proportion/multiplication predictions. Friedman's Test revealed all significant results for all conditions ($C_{0.00}$: $\chi^2(3) = 13.622$, $p = .003$; $C_{0.25}$: $\chi^2(3) = 130.795$, $p \leq .001$; $C_{0.50}$: $\chi^2(2) = 82.271$, $p \leq .001$; and $C_{0.75}$: $\chi^2(1) = 42.123$, $p \leq .001$) including the observed flat line of the non-contingent condition $C_{0.00}$. Similar to experiment 3.8, the significant results of conditions $C_{0.00}$ were attributed to the minority non-zero judgments that exhibit outcome density bias, of which captured by Friedman's Test. See Discussion of Experiment 3.1 for details.

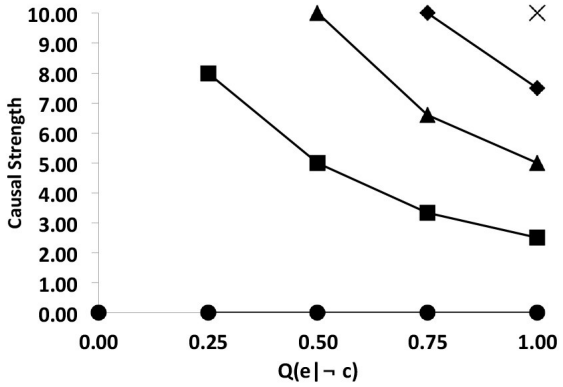


Figure 3.14: Results for Experiment 3.10. The left plot refers to consistent limit, while the right plot refers to higher limit condition.

Histogram analysis also showed that modes of all conditions reflected proportion/multiplication strategy. Tendency analysis on the data yielded a significant main effect of the scores after correction with Greenhouse-Geisser, $F(1.77, 106.23) = 69.62$, $p \leq .001$.

The mean scores for base rate neglect, proportion/multiplication, and difference were, respectively, $M = .31$, $SD = .02$; $M = .52$, $SD = .04$; and $M = .41$, $SD = .02$. Bonferroni *post hoc* test revealed significant pairwise differences between all of these scores: base rate neglect vs. proportion ($p \leq .001$), base rate neglect vs. difference ($p \leq .001$), and proportion vs. difference ($p \leq .001$).

3.10.3 Discussion

The results of this experiment indicated that participants adopted proportion/multiplication strategy the most, and difference strategy the least. This was consistent with the results of all previous preventive experiments (except in Experiment 3.9, where proportion/multiplication was the second mostly adopted, but difference strategy remained the least adopted). This cross-experiment consistency provides strong evidence of the relevance of the proportion/multiplication strategy in preventive scenarios. Between proportion and multiplication, until there is a way to properly disentangle them, I cannot test which strategy would be more dominant. Consequently, whether results of preventive reasoning are symmetrical with their generative counterpart remains a question. I will return to this point in General Discussion of Chapter 3 section.

3.11 General Discussion of Chapter 3

This chapter pursued the investigation to study causal reasoning with continuous outcomes, using a counterfactual judgment procedure. This procedure involved thinking of imagined intervention, which offers a purer way of tapping into causal reasoning procedure relative to the explicit judgment method as presented in Chapter 2.

In this chapter, six experiments involved generative causal relations, while four involved preventive relations. I adopted three techniques to analyse the results from these experiments: two techniques employed statistical hypothesis testing (i.e. trend, and tendency analysis), whereas the other involved qualitative observation (i.e. histogram analysis). The conventional trend analysis technique was less appropriate because of its use of central tendency measure. In many instances of the experiments, trend analysis produced significant results when qualitative inspection of the data clearly showed an absence of systematic trends. Thus, I did not fully rely on the results of this analysis, but focused on the other two instead. Refer to Figure 3.15 for a compilation of trend analysis results, Table 3.1 for tendency analysis results, and Figure 3.2 for histogram analysis results of all experiments in this chapter.

Accompanying the two proposed strategies for study (i.e. proportion, difference) was an unpredicted multiplication strategy. Although only two experiments in the generative scenario found evidence for this strategy, the results were consistent such that when participants had opportunity to reason using this multiplication strategy, judgments neglecting the base rates were reduced. In the next chapter, this strategy will be investigated from the very beginning, providing a richer insight onto its relevance for causal reasoning with continuous effects.

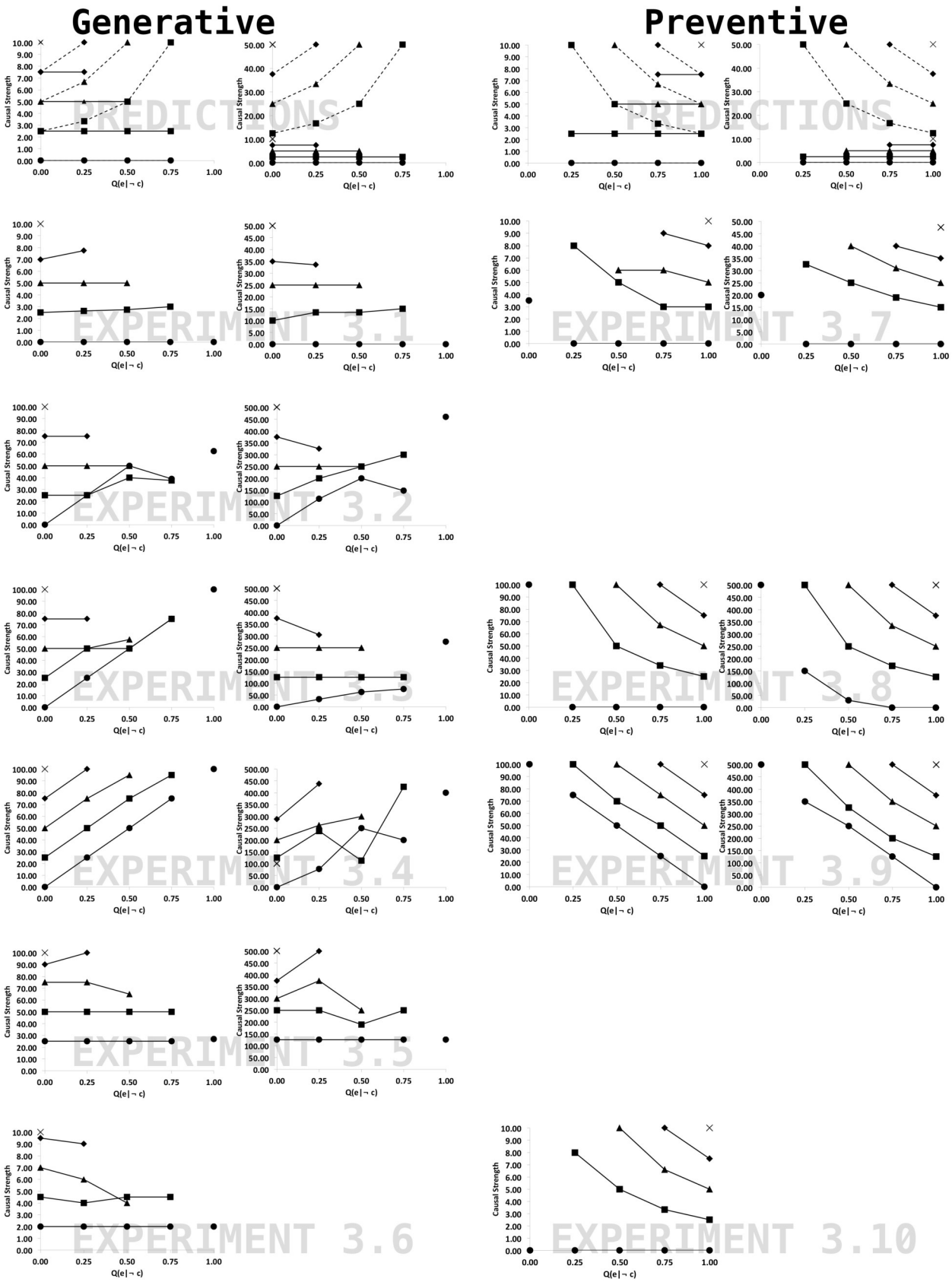


Figure 3.15: Compilation of results for all experiments in chapter 3. The left and right column consisted of generative and preventive experiments, respectively. For individual experiment the left plot refers to consistent limit, while the right plot refers to higher limit condition. The plots are arranged so that each row involves experiments in generative and preventive that have the same scenario.

As for the preventive scenario, the cross-experiment consistency provided strong evidence for the dominance of the proportion/multiplication strategy. Because their predictions were overlapping, I was unable to determine which of these two strategies were more prominent: whether the most adopted strategy in preventive and generative scenarios were consistent remains a question. Therefore, it is noteworthy to highlight potential contributors for the symmetry/asymmetry of both scenarios: it could be due to the different structure between generative and preventive scenarios such as limit saliency, different direction of influence between background causes and candidate cause. I discuss further on these factors in Chapter 5.

Table 3.1: Results of tendency analysis for all experiments

Generative Experiments (N)	Cover story		Means (*SD*) for Strategies			
	Theme	Design	Base rate Neglect	Proportion Strategy	Difference Strategy	Multiplication Strategy
1 (30)	Cream-Rash	W	.24(.02)	.40(03)	.69(.05)*	-
2 (30)	Fertilizer-Crop	B	.60(.05)*	.45(.03)	.40(.04)	-
3 (88)	Chemical-Algae	W	.53(.03)*	.42(.02)	.46(.03)	-
4 (18)	Chemical-Algae	B	.62(.09)*	.47(.05)	.44(.08)	-
5 (63)	Chemical-Algae	W	.37(.03)	.39(.03)	.54(.04)*	.50(.04)
6 (63)	Additive-Oil	W	.13(.01)	.38(.02)	.37(.02)	.59(.04)*
Preventive Experiments			Base rate Neglect	Proportion/ Multiplication Strategy	Difference Strategy	
7 (30)	Cream-Rash	W	.31(.02)	.52(.04)*	.41(.02)	
8 (89)	Chemical-Algae	W	.57(.03)	.58(.03)*	.42(.02)	
9 (17)	Chemical-Algae	B	.67(.08)*	.50(.06)	.33(.04)	
10 (61)	Additive-Oil	W	.31(.02)	.52(.02)*	.41(.02)	

Note: 'W' and 'B' in the third column refer 'within-entity' and 'between-entity' design as explained in the text. The tendency analysis aimed to capture each participant's tendency of using either one of the strategies using scores: the higher the score, the more incline the participant towards that strategy. The means in this table refer to across participants' average scores in each experiment.

Further, there was also a similar pattern across generative and preventive scenarios involving judgments of base rate neglect: These judgments were smaller in experiments using within-entity situation relative to between-entity. As per the elaboration in discussion of Experiment 3.2 and 3.3 above, having presenting the base rate information in a within-entity situation strengthen the link between this and the post-treatment information during the "undoing" as Sloman and Lagnado (2005) described in a counterfactual reasoning, relative to when in a between-entity situation. This pattern indicated that causal situation (i.e. between-entity versus within-entity) moderates the causal reasoning process involving continuous outcomes.

On another note, the counterfactual judgment approach involves two stages: the learning stage, and the applying stage. In the series of these experiments, I have assumed that participants were consistently using the same strategy in both stages. In other words, when participants read the cover story and received the contingency information, they utilised a certain strategy to bring about a causal index. Using this index, then the participants adopted the same strategy to answer the counterfactual question of which its answer reflects the adopted strategy. It was reasonable to assume that the strategies were consistent because the settings and context between the two stages were also consistent. Nonetheless, the results of these experiments indicated that context plays a bigger role than I initially thought; therefore, this assumption that people were consistent in both stages might not be fully warranted. An exploration of whether the strategies involved in these two stages are the same is noteworthy for future research.

Moreover, the findings from experiments with zero counterfactual base rate, where a majority of participants neglected the base rate, are consistent with what Perales and Shanks (2008) found. They reported that when adopting a counterfactual judgment approach, their

results indicated a relation with the probability of the ffect in the presence of the cause, P(e|c); in other words, participants were focusing on this value and ignored the base rate, resulting to what I reported as base rate neglect judgments. Perales and Shanks argued that it could be participants were confused about the wording of counterfactual questions. In their experiment of mutation and radiation, the question would be "*How many out of 100 butterflies, none of which would show a mutation in the test if unradiated, do you estimate would show a mutation if radiated?*" Instead of proper understanding, it seems easier for participants to understand that they need to imagine P(e|¬c) as zero, and then decide for P(e|c). Similarly, in this experiment, they could understand the question as asking them to set Q(e|¬c) to be zero before making judgment.

I pursued the exploration of causal reasoning with continuous outcomes in the next chapter using a method that was simpler and dependent of any assumption. Instead of explicitly asked for *causal strength* approximation in Chapter 2, or manifestation of the strength via counterfactual reasoning as in this chapter, the approach in Chapter 4 focuses on *causal direction* as the medium in tapping into strategy during causal reasoning.

Chapter 4: Empirical Investigation via Implicit Judgment

In Chapter 3 the investigation involved a counterfactual judgment approach. While this is a powerful method to study causal reasoning, it entails a naïve assumption that reasoners were consistently using the same strategy in both learning and applying stages.[10] To accompany this approach in exploring causal inference with continuous effects, I proceeded with another approach, namely implicit judgment.

This approach is modelled after Experiment 1 from Liljeholm and Cheng (2007). In their Experiment 1, they wanted to identify what aspect of causality people carries from one context to another (i.e. an invariant mental construct across contexts). The experiment did not require participants to make any explicit estimation of causal strength, but instead asked them to simply make a judgment about the existence of a simple causal relation. The experiment, however, involved probabilistic binary causation.

The cover story in the experiment was about two studies on the influence of two medicines (A and B) on headache. For each study, two groups of patients were involved: a treatment group, and a no-treatment (i.e. control) group. Participants were told that in a first study, a treatment group received only medicine A, whereas in a second study the treatment group received both medicine A and B. After having studied the datasets of both studies, participants simply had to give a 'Yes' / 'No' judgment of whether they thought medicine B causes headaches.

The experiment used a between subject design in which participants were divided into two groups. Both groups received the same cover story but different contingency conditions. One of the groups received a condition in which the causal power in study one, and the compound

[10] See beginning and General Discussion of Chapter 3 for more details of this assumption.

causal power (i.e. the net change from the control and the treatment group) in study two were at the same value. In the other group, the values of these causal powers varied. The former was referred to as power-constant group, while the later was referred to as ΔP-constant group. In both groups, even though the critical measure was their judgments on whether or not medicine B was a cause of headache, the psychological representation of interest was medicine A. This is because the judgment about medicine B reflects participants' assumption on the invariant aspect of medicine A across two different contexts (study one and two), as explained below.

I will use the mineral–algae context to illustrate the idea of the experiment.[11] From Chapter 1, take scenario 1 as study one, and scenario 2 as study two, but in study two, both mineral A and B were administered instead (see Figure 4.1 for an example). Using this example, therefore, in study one, 30 of the 100 pools already had algae growth even before receiving treatment with mineral A, and the number of pools covered with algae increased to 65 after receiving the treatment. Meanwhile in study two, none of the 100 pools had algae growth before receiving treatment with mineral A and B, and 50 of them had algae growth after receiving the treatment. Thus, in study one, the value of causal power was 0.50 and ΔP was 0.35, whereas in study two the compound value of causal power was also 0.50, but ΔP increased to 0.50. This is an example of a power constant problem. In this condition, therefore, when asked whether mineral B was a (another) cause of algae growth, participants who judged that it was not a cause showed sensitivity towards causal power as the invariance. This is because claiming that B is not a cause implies that all the causal change in study two must have occurred solely due to mineral A. In other words, a claim that B is non-causal reflects a belief that the net causal change (due to A) is identical across both scenarios, and thus a sensitivity to causal power (which was constant across both studies). In contrast, those who judged mineral B as a cause would demonstrate

sensitivity towards ΔP as the causal invariance: According to these participants, the net causal change in study two would have been higher than in study one, which means that A alone could not explain it, and B exerted a causal influence over and above that of A. Meanwhile, participants in the ΔP-constant group received the reverse condition: Across study one and two, ΔP remained constant, but the value of their causal powers varied (see Figure 4.1 for an example). Liljeholm and Cheng found that people's judgment patterns reflected sensitivity to causal power rather than ΔP as the invariant across contexts.

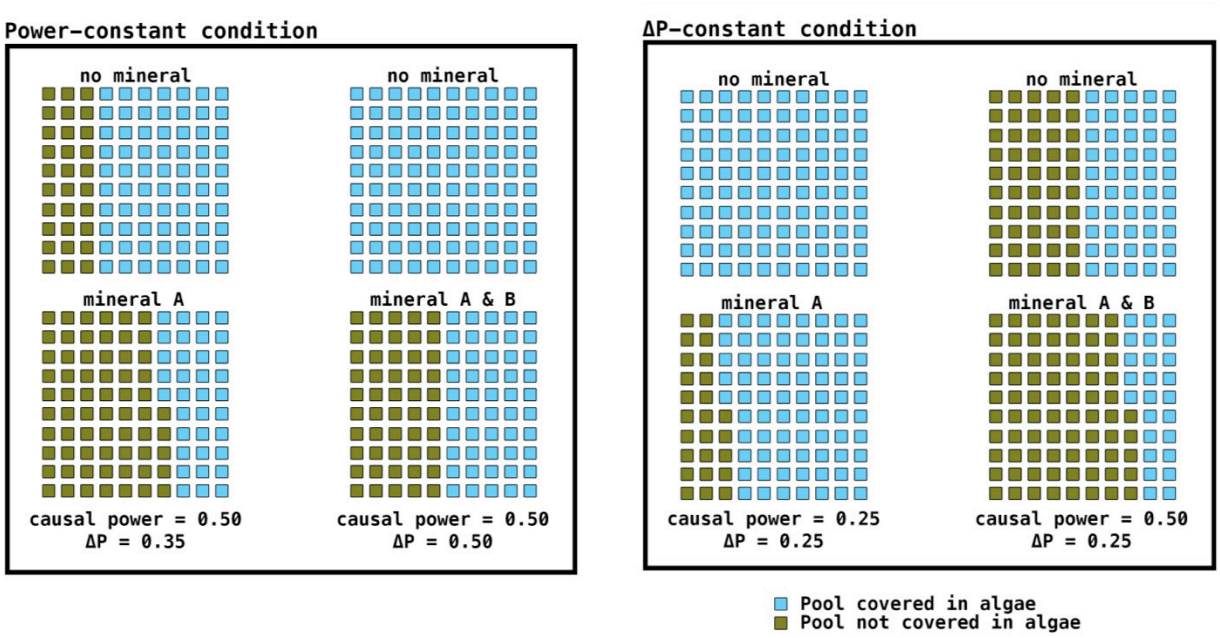


Figure 4.1. *An example of condition problem modelled after Experiment 1 in Liljholm and Cheng (2007). The top box portrays a power-constant condition while the bottom box portrays a ΔP-constant condition.*

Although it does not provide explicit information on causal strength, this implicit judgment approach has the advantage that it reduces biases that compromise other approaches like the counterfactual judgment approach as in Chapter 3, or explicit causal ratings on a scale as in Chapter 2, (see discussion in Buehner, Cheng, Clifford, 2003).

[11] Liljeholm and Cheng (2007) used a medicine-headache context in their study.

4.0 Experiment Overview

I adapted this design into the experiments to address the question on which strategy among the three (i.e. proportion, difference, and multiplication), reasoners would adopt when judging a causal relation involving a continuous outcome. In this chapter I report eight experiments; four of them used the same cover story about the influence of two minerals (causes) on the surface area of algae growth in a pool (continuous effect) at two different locations (contexts), while another four used a different cover story about the influence of two additives (causes) on runniness of engine oil (continuous effect was measured in terms of area of oil covering a test slate after being dripped from a certain height) at two different departments (contexts).

In these experiments, however, instead of asking participants *whether* or not mineral two, i.e. the mineral that exists only at one location, is a cause of algae growth, I asked *how* mineral two influences algae growth. The idea was to compare participants' 'judged direction of influence' (JDI) to the 'predicted direction of influence' (PDI) of each reasoning strategy to determine which strategy participants adopted. In the experiments, participants were given three JDI options: they could judge the direction of influence of mineral two on algae growth as either 'causing', 'inhibiting', or 'does not influence' algae growth.[12]

For example, consider this scenario 6 in which the area covered with algae at location one was $Q(e|\neg A) = 20\ m^2$ before treatment, and increased to $Q(e|A) = 80\ m^2$ after treatment with mineral A (see Figure 4.2 for an example of scenario 6). At location two, the area covered with algae before any treatment was $Q(e|\neg A,\neg B) = 30\ m^2$, and increased to $Q(e|A,B) = 90\ m^2$ after

treatment with minerals A and B. The surface area of pools at both locations was the same at 100 m^2. If three reasoners X, Y, and Z respectively judge mineral B in this example to inhibit, to cause, and to have no influence on algae growth, I would code this as JDI_X = 'inhibiting', JDI_Y = 'causing', and JDI_Z = 'does not influence'.

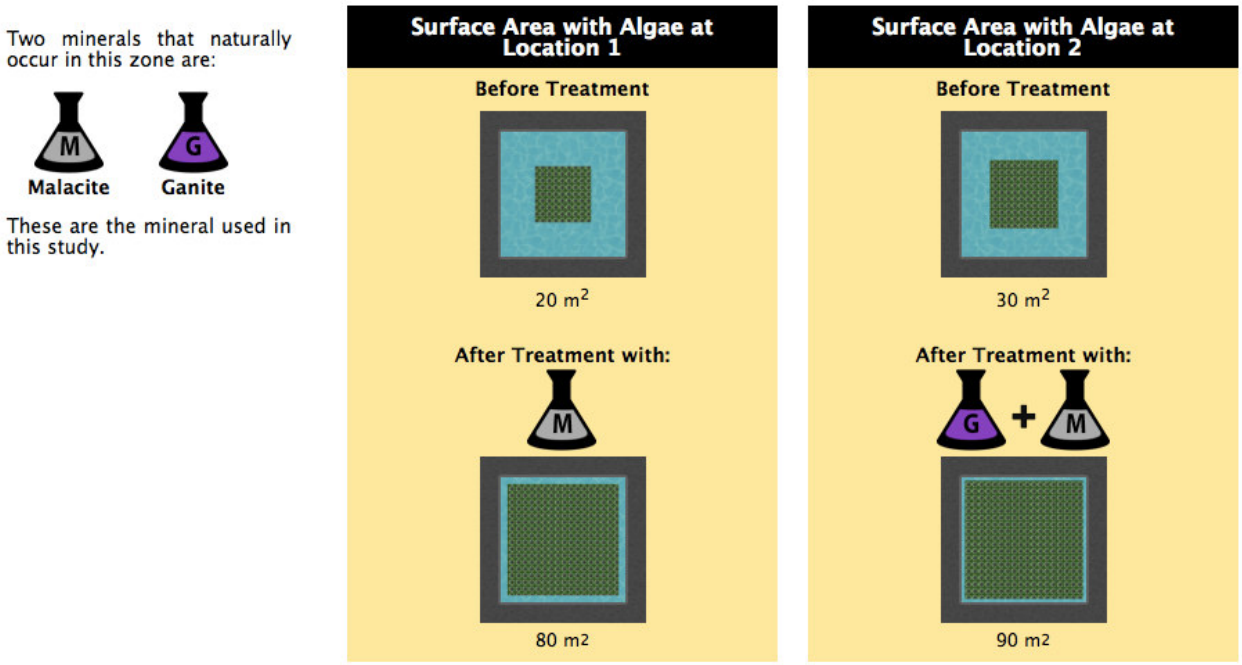


Figure 4.2. *A screenshot of condition [0.2, 0.8 : 0.3, 0.9]. This condition was also one of the conditions used in generative scenario of Experiment 4.1. In Experiment 4.2, a similar visual presentation was used but we modified it to reflect a between-entity experiment.*

To determine the PDI for any strategy, I computed the causal property of that strategy for both location one and location two, and then observed their direction of change between location one and location two. The direction (i.e. the PDI) could either be 'increasing', 'decreasing', or 'constant'. In the above scenario 6, the proportion index increases from 0.75 to 0.86 between locations ($\frac{80\ m^2-20\ m^2}{100\ m^2-20\ m^2} = 0.75; \frac{90\ m^2-30\ m^2}{100\ m^2-30\ m^2} = 0.86$), the multiplication index decreases from 4

[12] Some experiments have an additional fourth JDI option 'Cannot tell'. I introduced this option to avoid misunderstanding with the 'does not influence' option. This is described in more detail in the sections discussing

to 3 ($\frac{80\ m^2}{20\ m^2} = 4$; $\frac{90\ m^2}{30\ m^2} = 3$), and the difference index remains the same at 0.60 (80 m^2 – 20 m^2 = 60 m^2; 90 m^2 – 30 m^2 = 60 m^2). Hence, in this example, $PDI_{proportion}$ is 'increasing', $PDI_{multiplication}$ is 'decreasing', and $PDI_{difference}$ is 'constant'.

For generative scenarios, the PDIs map to JDIs as follows: A PDI of 'increasing' reflects that mineral 2 *causes* algae growth over and above the growth attributed to mineral 1; a PDI of 'decreasing' reflects that mineral 2 *inhibits* or *prevents* the algae growth that would occur due to mineral 1 alone; and a PDI of 'constant' reflects that mineral 2 *does not influence* algae growth over and above that attributed to mineral 1. In contrast, the mapping between PDI and JDI for the preventive scenario was reversed: A PDI of 'increasing' reflects that mineral 2 further *inhibits* or *prevents* the algae growth over and above inhibition attributed to mineral 1; a PDI of 'decreasing' reflects that mineral 2 *causes* algae growth that would otherwise be prevented by mineral 1; and a PDI of 'constant' reflects that mineral 2 *does not influence* algae growth over and above that attributed to mineral 1. Table 4.1 summarises these mappings. Because scenario 6 is an example of a generative scenario, from the mapping of each hypothetical reasoner's JDI with any of the PDIs, I can infer that reasoner X adopted a multiplication strategy, reasoner Y adopted a proportion strategy, and reasoner Z adopted a difference strategy.

Table 4.1: Mapping of JDIs with PDIs

PDIs	JDIs		
	causing	inhibiting	does not influence
increasing	*g*	*p*	
decreasing	*p*	*g*	
constant			*g p*

Note: The letters 'g' and 'p' represent mapping for generative and preventive scenario respectively.

each of the experiments.

4.0.1 Development of Conditions

The goal was to create conditions of two magnitude pairs that would produce distinct PDIs for each strategy. For these experiments, I achieved this by searching for pairs of magnitudes that would give distinct PDIs for each strategy. Table 4.2 lists the condition pairs I used for the generative scenarios, and Table 4.3 those for the preventive scenario. The following paragraphs describe the determination of these magnitude pairs.

Consider a [10 x 10] matrix of 100 magnitude pairs, each drawn from a range between 0.00 to 1.00 with an interval of 0.10, and representing Q(e|A) and Q(e|¬A).[13] For each of these 100 pairs, I could compute the causal properties of the three strategies and graph their relationships in a 3-dimensional scatter plot (see Figure 4.3). In Figure 4.3, the X-axis represents the index of the difference strategy, the Y-axis the proportion strategy, and the Z-axis (i.e. the colour) the multiplication strategy. The right half of the plot denotes the relationship among the three properties for generative causation, while the left half denotes preventive causation.

From these 100 pairs, I could search for two magnitude pairs that produced distinct PDIs for each strategy. Because any point in the space of the plot corresponded to one magnitude pair, searching for a suitable combination of two magnitude pairs meant identifying two points within either the left or right half of the graph (for preventive or generative). I could choose two initial points, which defined a line perpendicular to one of the axes. The intersection of the line with the axis denoted the property of the strategy for both points. Because both points (i.e. the two magnitude pairs) had the same property, they produced a 'constant' PDI of the strategy corresponding to the axis. For example, consider point 1 and 2 in Figure 4.3 that constructed line A that is perpendicular to the X-axis, which denotes the index of the difference strategy.

[13] The same principle applied for searching for the location two magnitude pair: Q(e|A,B) and Q(e|¬A,¬B).

Therefore, the $PDI_{difference}$ for a magnitude pair using points 1 and 2 is 'constant' because both points shared a difference index of of 0.60. I could then examine the other two strategies' PDIs for these two points to ensure they were pointing in opposite directions. If this was not satisfied, I could move the line along that axis and look for other possible locations ensuring that the other two PDIs are not identical. In the above example, $PDI_{proportion}$ increases from 0.75 to 0.86 when comparing points 1 and 2, while $PDI_{multiplication}$ decreases from 4 to 3. Consequently, the pair formed by points 1 and 2 satisfied the aim to have all PDIs pointing in different directions for the three strategies. Note that moving line A along the X-axis, always maintains $PDI_{proportion}$ and $PDI_{multiplication}$ pointing in opposite directions, while $PDI_{difference}$ remains constant.

I could use this procedure to determine other conditions. For the generative scenario, in addition to the above example, I could repeat the exercise but this time choosing two points to create a line perpendicular to the Y-axis (proportion index). In Figure 4.3, consider points 3 and 4 that created line B as an example. Using these two magnitudes, I obtain $PDI_{proportion}$ = 'constant' at 0.5, whereas $PDI_{difference}$ and $PDI_{multiplication}$ are both 'decreasing' (when moving from point 3 to point 4). It is evident in this mathematical representation that the nature of any pairs of points on any line perpendicular to the proportion-index-axis defines contrasts for $PDI_{difference}$ that are aligned with (i.e. pointing in the same direction as) $PDI_{multiplication}$. Points 5 and 6 in Figure 4.3 define line C that is perpendicular to the Z-axis (multiplication index). Using these two points, I obtain $PDI_{multiplication}$ = 'constant' at 2, but $PDI_{proportion}$ and $PDI_{difference}$ are both 'increasing' when considering point 5 as the starting point. As before, moving the line along the Z-axis would not disentangle the directional overlap between $PDI_{proportion}$ and $PDI_{difference}$.

In short, for generative causation, it is only possible to obtain distinct PDIs for the three strategies when the difference index is kept constant. In contrast, when the proportion index and

multiplication index are kept constant, it is not possible to obtain contrasting PDIs for the two remaining strategies. All points in the right side of Figure 4.3 represent conditions in Table 4.2 (i.e. the points 1 and 2 represent the condition of the first row from top, points 3 and 4 correspond to the condition in the third row, and points 5 and 6 correspond to the fourth row). Considering just these three conditions reveals that the PDI 'causes' is only represented for the proportion and difference, but not the multiplication strategy. Similarly, the option 'inhibits' is only represented for PDIs of the difference and multiplication strategies, and not for the proportion strategy. To address this I included the condition in the second row of Table 4.2. In sum, I created conditions so that each of the three reasoning strategies is represented at least once for each PDI, and doing this meant that for some of the conditions two strategies overlapped in their PDI, which in turn meant that in these conditions one PDI did not map to any of the three strategies considered here.

To identify conditions for the preventive scenario, I used the same procedure. Consider points 11 and 9 on line B in Figure 4.3 for an example of the first search on preventive magnitude pairs. Both share $PDI_{proportion} = 0.50$; however, unlike their generative counterpart (defined by points 3 and 4), they also share an identical $PDI_{multiplication} = 2$, while $PDI_{difference}$ is 'decreasing' from -.3 to -.5. If I fixed $PDI_{difference}$ as 'constant' at -0.6 by choosing points 7 and 8 on line D, I would analogously find that both $PDI_{proportion}$, and $PDI_{multiplication}$ were pointing in the same direction, which was 'decreasing' from 0.75 to 0.67, and from 4 to 3, respectively. In other words, in this mathematical setup, the nature of the relationship among the strategies in preventive causation always gives an overlapping PDI for multiplication and proportion strategies. Therefore, preventive scenarios can only distinguish difference strategies versus proportion/multiplication. Similarly to the generative scenario, to complete the conditions so that

all strategies' predictions are present in all judgment options, I included another condition (in Figure 4.3, this additional condition is represented by points 9 and 10).

4.0a Notation

I use the following notation when referring to any condition in all experiments of Chapter 4: condition [$Q(e|\neg c)_1$, $Q(e|c)_1$: $Q(e|\neg c)_2$, $Q(e|c)_2$]. The first two items with subscript 1 before the colon represent the surface area covered with algae at location 1, whereas the next two items with subscript 2 represent algae growth area at location 2. As an example, I can simply refer to the condition in scenario 6 as [0.2, 0.8 : 0.3, 0.9].

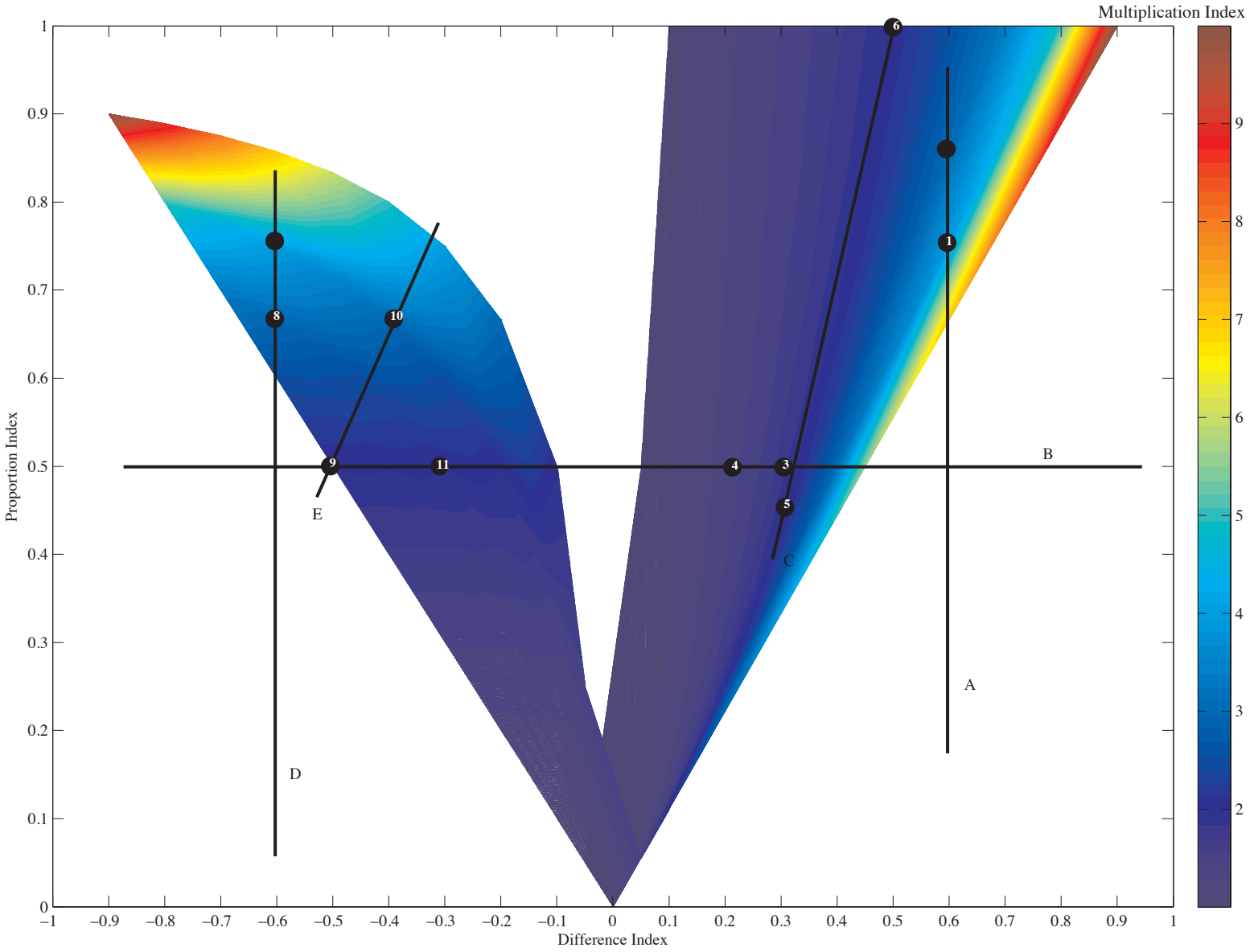


Figure 4.3. *Scatter plot of the predictions from the three strategies (proportion, difference, multiplication) using 100 sets of magnitude pairs from 10 different values of Q(e|c) and Q(e|¬c). The left half of the plot corresponds to relationship the three predictions have in preventive scenarios, while the right half corresponds to the relationship in generative scenarios. The X-axis denotes the difference strategy index, the Y-axis denotes the proportion strategy index c, and the Z-axis (i.e. the color) denotes the multiplication strategy index. All of the numbered points and the lines connecting them, refer to points in an example to describe the procedure to find conditions for the experiments.*

Table 4.2: *Experiment Conditions and Results of Experiment 4.1, 4.2, 4.3, and 4.4.*

Magnitude Pair at Location 1 (%)		Magnitude Pair at Location 2 (%)		Absolute Property	How does *cause 2* influence *effect magnitude*? *Results:* $N(\%)_{Exp.4.1}$ $N(\%)_{Exp.4.2}$ $N(\%)_{Exp.4.3}$ $N(\%)_{Exp.4.4}$				Total N Exp. 4.1 Exp. 4.2 Exp. 4.3 Exp. 4.4
Q(e|~c)	Q(e|c)	Q(e|~c)	Q(e|c)		Causes	Inhibits	Does not influence	'Cannot tell'	
0.20	0.80	0.30	0.90	P: 0.75 → 0.86	**P**	**M**	**D**		
				D: 0.60 → 0.60	*8(20.0)*	*13(32.5)*	*19(47.5)*	--	40
				M: 4 → 3	*32(53.3)*	*4(6.7)*	*20(33.3)*	*4(6.7)*	60
					8(23.5)	*1(2.9)*	*23(67.6)*	*2(5.9)*	34
					32(53.3)	*15(25.0)*	*12(20.0)*	*1(1.7)*	60
0.50	1.00	0.20	0.60	P: 1.00 → 0.50	**M**	**D/P**	**?**		
				D: 0.50 → 0.40	*17(42.5)*	*14(35.0)*	*9(22.5)*	--	40
				M: 2 → 3	*16(26.7)*	*41(68.3)*	*2(3.3)*	*1(1.7)*	60
					5(14.7)	*22(64.7)*	*6(17.6)*	*1(2.9)*	34
					18(30.0)	*32(53.3)*	*8(13.4)*	*2(3.3)*	60
0.40	0.70	0.60	0.80	P: 0.50 → 0.50	**?**	**D/M**	**P**		
				D: 0.30 → 0.20	*9(22.5)*	*17(42.5)*	*14(35.0)*	--	40
				M: 1.75 → 1.33	*18(30.0)*	*15(25.0)*	*25(41.7)*	*2(3.3)*	60
					3(8.8)	*12(35.3)*	*17(50.0)*	*2(5.9)*	34
					25(41.7)	*14(23.3)*	*19(31.7)*	*2(3.3)*	60
0.30	0.60	0.50	1.00	P: 0.43 → 1.00	**D/P**	**?**	**M**		
				D: 0.30 → 0.50	*16(40.0)*	*1(2.5)*	*23(57.5)*	--	40
				M: 2 → 2	*51(85.0)*	*0(0.0)*	*7(11.7)*	*2(3.3)*	60
					25(73.5)	*5(14.7)*	*3(8.8)*	*1(2.9)*	34
					39(65.0)	*5(8.3)*	*12(20.0)*	*4(6.7)*	60

Note: *The italicised numbers in the first, second, third, and fourth line in each row are the results of Experiment 4.1, 4.2, 4.3, and 4.4 respectively. The numbers are frequencies of judgments with their percentages inside the parenthesis. Legend: P, M, D denotes the three strategies, 'proportion', 'multiplication', and 'difference' respectively.*

Table 4.3: *Experiment Conditions and Results of Experiment 4.5, 4.6, 4.7, and 4.8.*

Magnitude Pair at Location 1		Magnitude Pair at Location 2		Absolute Property	How does mineral 2 influence algae growth? **Reasoning Strategy** *Results: n(%)$_{Exp.4.5}$ n(%)$_{Exp.4.6}$ n(%)$_{Exp.4.7}$ n(%)$_{Exp.4.8}$*				Total N Exp. 4.5 Exp. 4.6 Exp. 4.7 Exp. 4.8
Q(e|~c)	Q(e|c)	Q(e|~c)	Q(e|c)		causes	inhibits	does not influence	'Cannot tell'	
0.80	0.20	0.90	0.30	P: 0.75 → 0.67	**P/M**	**?**	**D**		
				D: 0.60 → 0.60	*12(34.3)*	*10(28.6)*	*13(37.1)*	*0(0.0)*	*35*
				M: 1/4 → 1/3	*23(38.8)*	*11(18.3)*	*22(36.7)*	*4(6.7)*	*60*
					21(35.0)	*12(20.0)*	*19(31.7)*	*8(13.3)*	*60*
					31(51.7)	*13(21.7)*	*13(21.7)*	*2(3.3)*	*60*
1.00	0.50	0.60	0.20	P: 0.50 → 0.67	**D**	**P/M**	**?**		
				D: 0.50 → 0.40	*5(14.3)*	*27(77.1)*	*3(8.6)*	*0(0.0)*	*35*
				M: 1/2 → 1/3	*4(6.7)*	*48(80.0)*	*0(0.0)*	*8(13.3)*	*60*
					3(5.0)	*45(75.0)*	*5(8.3)*	*7(11.7)*	*60*
					7(11.7)	*46(76.7)*	*6(10.0)*	*1(1.7)*	*60*
0.60	0.30	1.00	0.50	P: 0.50 → 0.50	**?**	**D**	**P/M**		
				D: 0.30 → 0.50	*9(25.7)*	*8(22.9)*	*18(51.4)*	*0(0.0)*	*35*
				M: 1/2 → 1/2	*19(31.7)*	*15(25.0)*	*21(35.0)*	*5(8.3)*	*60*
					19(31.7)	*17(28.3)*	*19(31.7)*	*5(8.3)*	*60*
					22(36.7)	*19(31.7)*	*18(30.0)*	*1(1.7)*	*60*

Note*: The italicised numbers in the first, second, third, and fourth line in each row are the results of Experiment 4.5, 4.6, 4.7, and 4.8. The numbers are frequencies of judgments with their percentages inside the parenthesis. Legend: P, M, D denotes the three strategies, 'proportion', 'multiplication', and 'difference' respectively.*

Experiment 4.1

Experiment 4.1 examined reasoning about continuous outcomes in a generative scenario, using a cover story on algae growth similar to the examples used above. Participants were informed that research was conducted in four different climatic zones (to motivate the four different conditions listed in Table 4.2), and that each condition involved evaluating the results of research conducted at two locations situated within the same climatic zone. Among the four conditions (see Table 4.2), three of them had overlapping PDIs. An overlap of PDIs logically implies that one of the options of JDIs available to participants (i.e. 'causes', 'inhibits', 'does not influence') does not match with any of the strategies I was investigating. These cases are indicated with a question mark in Table 4.2.

4.1.1 Method

Participants. Fourty undergraduate students from the School of Psychology, Cardiff University participated to fulfil part of a course requirement. Participants for this experiment were those who had not participated in any of previous experiments.

Design and Procedure. Each participant worked on the four conditions listed in Table 4.2, presented in a random order. The experiment was conducted over the Internet. Participants began by reading a cover story about research on the influence of minerals on algae growth in four different climate zones: tropical, arid, mediterranean, and alpine, and were told that for each zone, two naturally existing minerals were selected and studied at two locations. In location 1, only mineral one was examined, whereas in location 2 both minerals were examined. Eight different fictitious names of minerals were used and randomly assigned to conditions and roles (mineral 1 and 2). Instructions also stated that due to both locations being many miles apart, their microclimates might be different. This was done to motivate differences in base rate (i.e.

coverage with algae in the absence of treatment) between the two locations. Participants then read that examination of minerals took place in pools with a surface area of 100 square meters, and that the area of water surface covered by algae was used as the marker of algae growth. After reading the cover story, participants proceeded with the conditions.

Each participant worked on the four conditions listed in Table 4.2, with every condition randomly assigned to one climatic zone, and order of conditions randomised. For each condition, I presented all related information simultaneously on a single screen (see Figure 4.2). This included a visual representation of algae growth before and after treatment with the minerals at both locations: information for location 1 on the left side, and for location 2 on the right side. The growth area reflected the effect magnitudes represented in Table 4.2 with respect to 100 m^2 of the pool surface area.

Underneath this information was a question: *"Based on the information from BOTH locations, how would you judge the influence of [mineral two] on algae growth?"* Participants had three options to choose whether mineral two 'causes', or 'does not influence', or 'inhibits' the growth of algae. After submitting the judgment, they received the next condition in the same visual format. For complete instructions see Appendix 1.

4.1.2 Results

The right side of Table 4.2 displays the results of Experiment 4.1. Qualitative observation suggests that multiplication is the most prominent strategy, followed by difference. Proportion, on the other hand, appears to be the least adopted strategy. While I cannot make a clear distinction between multiplication and difference in condition [0.4, 0.7 : 0.6, 0.8], I can observe that the multiplication strategy is the most chosen in conditions [0.5, 1.0 : 0.2, 0.6] and [0.3, 0.6 : 0.5, 1.0]. In condition [0.2, 0.8 : 0.3, 0.9], the difference strategy is the most prominent.

Perfect strategy consistency within participants across conditions was low: No participant consistently adopted the proportion strategy across all four conditions. Meanwhile, only five and two participants consistently adopted the multiplication and difference strategy, respectively. Therefore, I analysed the extent of consistency (instead of perfect consistency) of judgments across conditions for every participant using scores. To this end, I converted the four judgments each participant made into three scores corresponding to each strategy. Any judgment (i.e JDI) that corresponded to a particular PDI of a given strategy (including in an overlapping PDI situation) contributed one point towards the total score of that strategy. For example, a participant who answered 'does not influence', 'inhibits', 'inhibits', 'causes' respectively to the four conditions in the order of Table 4.2 accrued a total of two points for proportion through judgments of condition [0.5, 1.0 : 0.2, 0.6] and [0.3, 0.6 : 0.5, 1.0], a total of four points for difference through all conditions, and one point for multiplication through condition [0.4, 0.7 : 0.6, 0.8].

The vertical-stripe columns in the top panel of Figure 4.4 show the results of this scoring analysis. The multiplication strategy attracted the highest score followed by the difference strategy. Meanwhile, the proportion strategy earned the lowest score. I conducted a Wilcoxon Signed-Ranks Test to compare the scores against the value expected by chance. Given that participants had three response options, a particular JDI therefore had a 0.33 chance of being selected on a given condition (including those with overlapping JDIs). Consequently, the overall chance score for a particular strategy would have been 1.33 for each participant (0.33 times four conditions). I compared the three strategies separately with individual Wilcoxon-Signed-Ranks tests against the chance value. Throughout the remainder of this chapter, I report the original *p* values (i.e. for single tests) but my evaluation is based on Bonferroni corrected thresholds.

Consequently, for an individual comparison to be significant at the .05 level, the *p* value would need to be below .0167. The difference strategy (z = -2.005, p = .045) and the multiplication strategy scores (z = -1.974, p = .048) were marginally above chance when considered as individual comparisons, but failed to pass the Bonferroni corrected threshold. The proportional strategy score was not significantly different from chance (z = -.176, p = .860).

4.1.3 Discussion

The scoring analysis further supports my initial observation that participants mostly adopted a multiplication strategy, and avoided a proportion strategy. This suggest that participants thought of the cause as interacting with the background causes, instead of considering the candidate cause as directly changing the effect. In other words, they might have conceived of algae growing on their own (due to the background causes), with the candidate cause merely amplifying this tendency, rather than acting as a cause on its own. In this case, the candidate cause influences the propensity of the background causes to produce the effect. This violates the normative assumptions of a proportional framework, namely that the influence of background causes on the effect should remain constant before and after treatment, and that the cause and background each influence the effect independently. Perhaps in the algae-mineral cover story this assumption was not salient enough.

In addition to the cover story, the within-entity reasoning situation of this experiment, whereby algae coverage changed from before and after treatment within the same pool, also permitted participants to consider that the causal efficacy of background causes changed as a function of the mineral. This is because the before-after change entails a change that happens *over time*. Experiment 4.1 contained no observable reference showing the magnitude of the effect in the absence of treatment, *but at the same later time.* Consequently, this may have triggered

uncertainties about whether the influence of background causes on the effect remained constant throughout the study or changed as a function of mineral administration (for more discussion on this, see Experiment 4.3).

To continue with the investigation, I considered two possible factors that might challenge the generalizability of results of this experiment: For Experiment 4.2, I developed a different cover story with the aim to convey that the causal influence of background causes remained constant throughout; in Experiment 4.3, I presented the information in a between-entity situation.

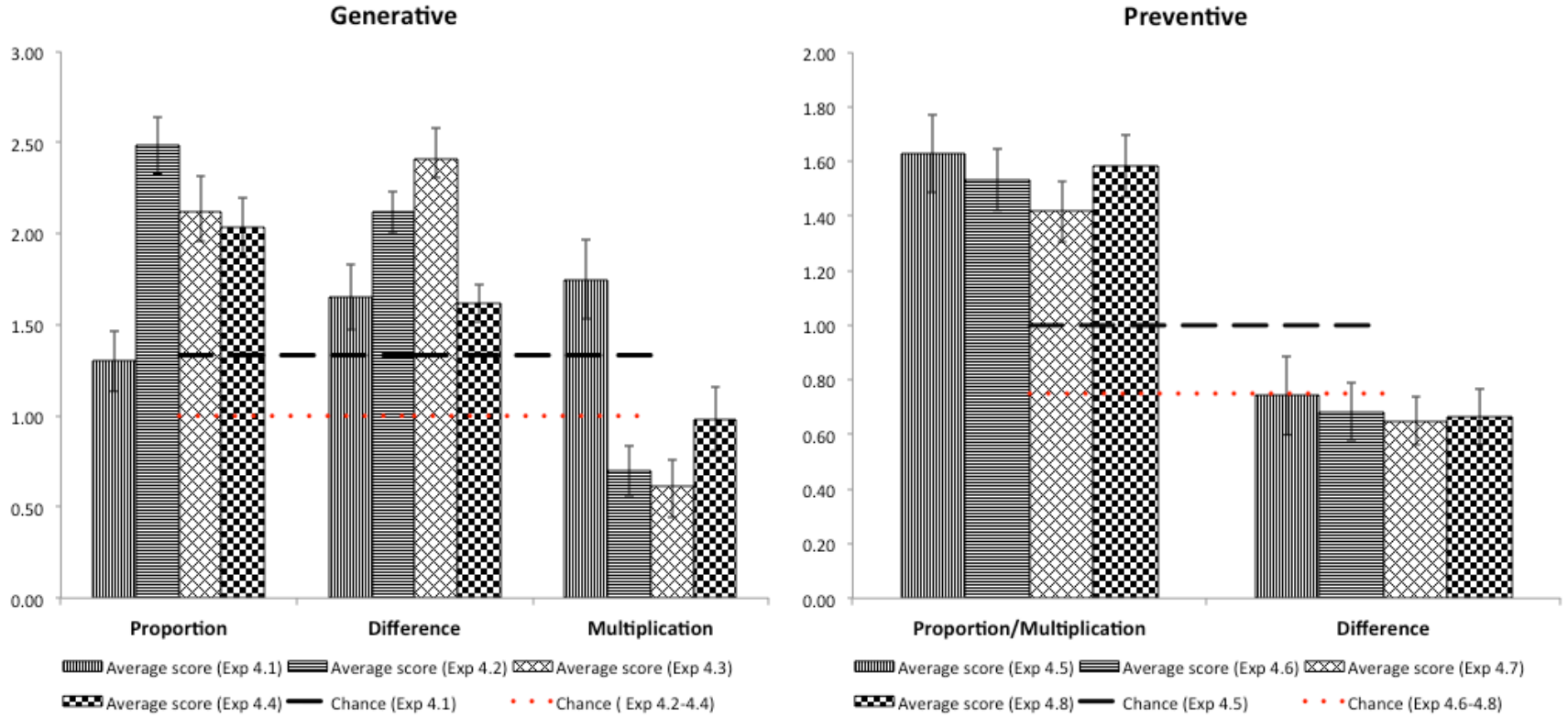


Figure 4.4: *Results of the scoring analysis for all experiments in Chapter 4. The top chart is for the generative, the bottom for the preventive scenarios. The dashed and dotted lines respectively represent chance values in the corresponding experiments.*

Experiment 4.2

In Experiment 4.1, the cover story of minerals influencing algae growth could plausibly have biased participants into thinking that the background causes and candidate cause interact to influence algae growth. More specifically, participants could have thought that algae growth was occurring due to background causes, and that the candidate cause was merely amplifying or modulating this process instead of acting directly on the effect. In this experiment, I used a new cover story about the influence of chemical additives (cause) on runniness of engine oil (effect), measured in terms of splash area covered by the oil before and after treatment. The rationale of doing this was that naïve assumptions about the causal mechanisms influencing viscosity of oil would be qualitatively different to those concerning algae growth. Specifically, I hoped that this new scenario would increase the likelihood that participants view the candidate causes (i.e. additives) as directly influencing the runniness of oil, rather than via interacting with the background conditions, or modulating a naturally occurring process.

4.2.1 Method

Participants. Sixty participants were recruited via Amazon Mechanical Turk and were paid USD 0.80. I restricted participation to the United States, and native English speakers only.

Design and Procedure. This experiment adopted the same within-subject design as Experiment 4.1 except with a different cover story. Participants began by reading a cover story about a study of two additives having influence on runniness of four engine oils (i.e. to correspond to four conditions). Participants were told that runniness was measured by a Drip-Test procedure: 5 grams oil would be dropped from 5 centimetres height onto a test slate of 10 square centimetres, and the spread (surface area covered in oil) would be determined. Participants were told that each oil went through an initial test, followed by the mixing with

additives, before being tested again. The context of the cover story involved two departments, with Department A testing only one additive while Department B tested both additives. The story also made it clear that "*The greater the surface area on the test slate (out of the 10 cm^2 total) covered by oil, the greater the runniness of the oil*".

After reading the cover story, participants continued to the first problem (randomly selected among the four). For each condition, participants were exposed to a visual presentation of the splash area before and after treatment for both departments (this was analogous to Figure 4.2, except that the visuals were slightly modify to depict a round slate and a blob of varying size). Underneath, "*Based on the information from BOTH departments, how would you judge the influence of [additive two] on the runniness of oil?*" Participants had four options to choose whether additive two 'causes', or 'does not influence', or 'inhibits' the runniness of oil, or 'Cannot tell'. I added this additional answer option of 'Cannot tell' to reduce any potential confounding participants might have had between a 'does not influence' judgment and their inability to make an actual judgment. If they chose this fourth option, they had to type in a reason as for why they could not make any judgment. I included this requirement as a deterrent from using this option as a lazy alternative to proper engagement with the task. Consequently, I did not systematically analyse answers to this question.

4.2.2 Results

The right side of Table 4.2 outlines the results of this experiment and suggests that the proportion strategy was the most adopted strategy in this experiment. Proportion-based judgments dominated condition [0.2, 0.8 : 0.3,0.9] and [0.4, 0.7 : 0.6, 0.8], while the overlapping of proportion and difference-based judgments dominated the other two conditions. Perfect strategy consistency within participant across conditions was relatively low (19 out of 60

participants) but better than Experiment 4.1. Out of the consistent 19 participants, 3 adopted the multiplication strategy, 4 the difference strategy, and 12 the proportion strategy. The horizontal-stripe columns in the top panel of Figure 4.4 represent results from the same scoring analysis as in Experiment 4.1.[14] Wilcoxon Signed-Ranks Tests on the judgments indicate that the proportion (z = -5.994, p < .001) and difference scores (z = -5.978, p < . 001) were both significantly above what would be expected by chance, whereas the multiplication score was significantly below chance (z = -2.673, p = .008).

4.2.3 Discussion

The initial observation that the proportion strategy was the most prominent was supported in the scoring analysis. These results are at variance with Experiment 4.1, where proportion was the least adopted strategy. Because the mathematical structure of the problems was identical between these experiments, and both presented participants with causal changes within the same entities (i.e. before and after treatment), the inconsistency between these two experiments must be resulting from the difference in cover stories. Specifically, I propose that the different cover stories engendered different assumptions about the causal processes involved in the scenarios: The algae-mineral story involved background causes that were naturally occurring in the environment (e.g. algae growth due to exposure to sunlight, nutrients, microorganisms, etc), which would continue to keep producing algae growth over time. Specifically, the causal power of these background causes would be expected to change over time (e.g. as the amount of microorganisms in a body of water grows, the potential to facilitate algal bloom also rises in a commensurate manner). Consequently, a natural assumption in this scenario indeed is that any treatment added to the pools would act on these background causes (i.e. a causal interaction),

[14] The chance level was 1.00 for every participant (i.e. 0.250 per condition, times four conditions). It was different from 1.333 in Experiment 4.1 because in this experiment, participants had four answer options per

rather than acting independently. In contrast, the oil-additive story entailed background causes that would be expected to remain constant across time (e.g. the extent to which ambient temperature, humidity etc. influence the spread of oil). Consequently, participants might have approached these problems with the assumptions that any influence these background causes have on the runniness of oil would remain constant between the before- and after treatment tests, and furthermore, that the additives act independently, rather than by influencing the background factors.

Experiment 4.3

In this experiment, I wanted to test whether the reasoning situation (within-entity vs between-entity) moderates the reasoning strategy people adopted. To this end, I reverted to the cover story used in Experiment 4.1, but modified it to a between-entity test of causal efficacy. Specifically, rather than measuring the surface area covered in algae before and after treatment, in this experiment, participants were introduced to two separate pools in each laboratory, one which served as a control pool and was left untreated, and another which received treatment with the mineral(s).

4.3.1 Method

Participants. Thirty-five undergraduates students from the School of Psychology, Cardiff University participated in the experiment to fulfil part of a course requirement.

Design and Procedure. Design and procedure were identical to Experiment 4.1, apart from small modifications to the cover story to reflect the between-entity situation. Specifically, I mentioned that each laboratory housed two pools: one pool served as a control, and received no treatment, whereas the other pool received treatment with either one (laboratory at location 1) or both (location 2) minerals. To further clarify the cover story, I included a diagram to show the

---

condition, relative to three in Experiment 4.1.

between-entity situation (Figure 4.5). As in Experiment 4.2, the fourth additional answer option of 'Cannot tell' was also available.

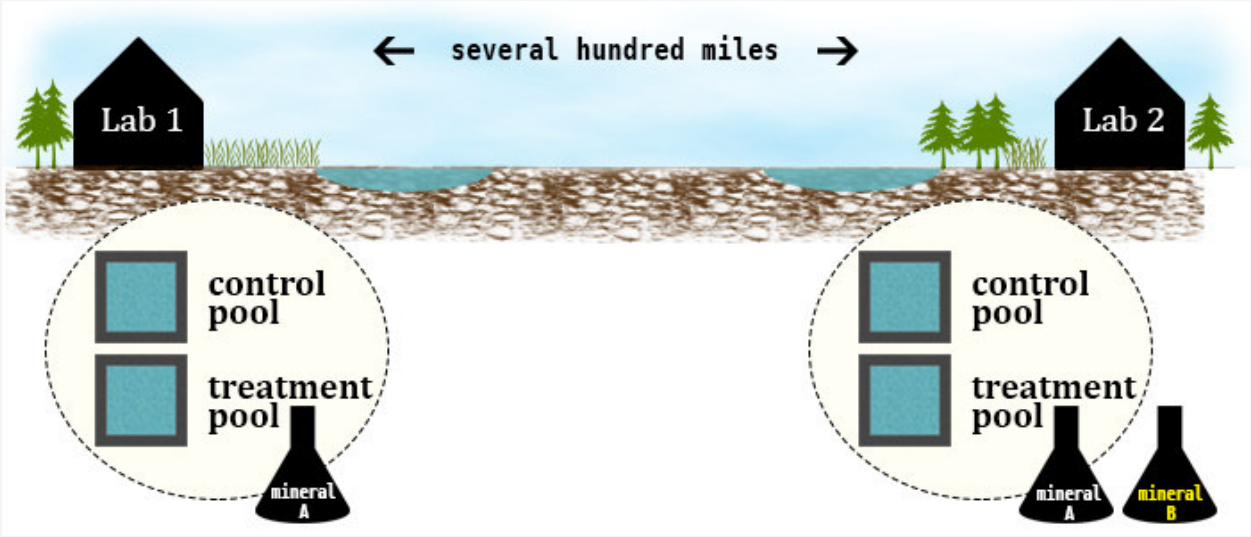


Figure 4.5. *A diagram included inline of the cover story to explain the set up of the situation in Experiment 4.3 and 4.7.*

4.3.2 Results

The right side of Table 4.2 summarises the results of the experiment. While the results are less clear than in the previous experiments, it is evident that the multiplication strategy was the least selected option in all conditions. In condition [0.5, 1.0 : 0.2, 0.6] and [0.3, 0.6 : 0.5, 1.0] most participants made judgments corresponding to the overlapping proportion/difference strategy. Meanwhile, in condition [0.2, 0.8 : 0.3, 0.9], the predominant choice was for the difference strategy, but in condition [0.4, 0.7 : 0.6, 0.8], the most common strategy was proportion. Similar to Experiment 4.1, perfect strategy consistency within participant across conditions was very low. Across all conditions, the number of participants who consistently used the multiplication, proportion, and difference strategies was respectively zero, two, and three. Therefore, I also conducted the same scoring analysis on these judgments to provide an indication of the most popular strategy. The cross-hatched columns in the top panel of Figure 4.4

represent the outcome of this analysis. Wilcoxon Signed-Ranks Tests on the judgments indicate that the proportion (z = -4.111, p < . 001) and difference scores (z = -4.682, p < . 001) were both significantly above chance, whereas the multiplication score (z = -2.502, p = .012) was significantly below chance.

4.3.3 Discussion

My initial observation was that the multiplication strategy was the least adopted strategy, and the scoring analysis further supported this. The majority of judgments in condition [0.5, 1.0 : 0.2, 0.6] and [0.3, 0.6 : 0.5, 1.0] corresponded to the overlapping of difference and proportion strategies. Thus, to determine which strategy of these two was dominant, only results in conditions [0.2, 0.8 : 0.3, 0.9] and [0.4, 0.7 : 0.6, 0.8] are relevant. The results in these conditions however, are not in agreement, as participants mostly adopted the difference strategy in the former and the proportion strategy in the latter. Thus, in this experiment it is inconclusive whether people adopted the proportion or the difference strategy.

Comparing these results to the dominance of the multiplication strategy in Experiment 4.1 suggests that reasoning situation (within-entity vs. between-entity) moderates the choice of reasoning strategy. In the within-entity situation of Experiment 4.1, because the treatment happened on the same entity, the credibility of information about effect magnitude in the absence of the candidate cause (i.e. only in the presence of background causes) was reduced. In other words, reasoners could not be certain that the influence of background causes would remain constant following the treatment. Because of this, reasoning about an interaction between background causes and candidate cause might have been more prominent, which could have led participants to adopt a multiplication strategy. In contrast, in the between-entity situation of Experiment 4.3, the effect magnitude in the absence of the candidate cause was observable and

verifiable for the control condition, even after the administration of treatment. This gives certainty to the constancy of the influence of background causes throughout the study. Because of this, thinking about an interaction may have been less likely, which could have contributed to multiplication being the least adopted strategy. Indeed, the fact that many judgments reflected use of the proportion strategy suggests that between-entity situations further strengthen the assumption of independence between background causes and candidate cause.

Experiment 4.4

The contrasting results of Experiment 4.1 and 4.3 suggest that reasoning situation (within-entity vs. between-entity) moderates the choice of reasoning strategy: multiplication strategy in the former, and proportion/difference strategy in the latter. These experiments, however, used the algae-mineral cover story. A natural question following these results would be whether reasoning-situation also moderates strategy choice in the oil-additive story. Because Experiment 4.2 tested a within-entity situation, I replicated the experiment here with a between-entity situation. This allowed us to investigate whether the moderation of strategy choice by reasoning-situation is generic or specific to the algae-mineral story.

4.4.1 Method

Participants. Sixty participants were recruited via Amazon Mechanical Turk and were paid USD 0.80. Participation criteria was the same as in Experiment 4.2.

Design and Procedure. Design and procedure were identical to Experiment 4.2, with the cover story modified to a between-entity situation. As in Experiment 4.3, the cover story referred to two separate observations in each location: One of a sample of engine oil without any additives added, and another where one or both additives were present.

4.4.2 Results

The results of this experiment occupy the right side of Table 4.2. A visual inspection suggests that proportion was the leading strategy. The proportion strategy was dominant in condition [0.2, 0.8 : 0.3, 0.9], and conditions [0.5, 1.0 : 0.2, 0.6] and [0.3, 0.6 : 0.5, 1.0] elicited the overlapping proportion/difference option as most frequent judgments. In condition [0.4, 0.7 : 0.6, 0.8], the most chosen option corresponded to the unknown PDI marked with the question mark followed by the proportion strategy.

Results of this experiment also have a low perfect strategy consistency within participants across conditions. Only one, five, and seven participants consistently adopted the difference, proportion, and multiplication strategies in all conditions, respectively. Results of the same scoring analysis as in the previous experiments are depicted in the chequered columns in the top panel of Figure 4.4. The highest score was associated with the proportion strategy followed by the difference strategy and the multiplication strategy warranted the lowest score. A Wilcoxon Signed-Ranks Test showed that scores for the proportion (z = -5.110, p < . 001) and difference strategy (z = -4.750, p < . 001) were significantly above chance, whereas the score for the multiplication strategy was not significantly different from chance (z = -.676, p = .499).

4.4.3 Discussion

Proportion was the prominent strategy in this experiment. Even though the reasoning situation in this experiment (between-entity) was different to Experiment 4.2 (within-entity), the results converge across both experiments. This is at variance with the pattern of results from Experiment 4.1 vs. 4.3, where the reasoning situation likewise varied, but the scenario (cover story) remained constant. I attribute the difference between Experiment 4.1 vs. 4.3, versus Experiment 4.2 vs. 4.4 to differences in the cover story between these two sets of experiments,

and the causal mechanisms they suggest (i.e. algae-mineral in Experiment 4.1 and 4.3, and oil-additive in Experiment 4.2 and 4.4). In short, the impact of reasoning-situation, which was present in the algae-mineral scenario, was not present in for oil-additive scenario.

A possible explanation for the convergence of the proportion strategy in Experiment 4.2 (within-entity) and this experiment (between-entity) was due to the clarity of the reasoning situation across these experiments. The sequence of events in the *within-entity* story began with dripping of oil for a base-rate measurement, followed by administering treatments on the remaining oil, and then dripping it for a subsequent measurement. The *between-entity* story began with preparation of two distinct samples right from the start, whereby the treatment(s) were administered only to one sample, and then both samples were dripped for measurement of the splash area. The subtle difference between these two stories was only the time when the measurement of base-rate oil and treated oil was carried out, but the end product was still two physically distinct splashes in both situations. In the algae-mineral story, only the between-entity version involved two physical pools, while in the within-entity version, the base-rate area was not physically observable once the treatment was administered. This was highlighted in Experiment 4.3: the difference between the within-entity and between-entity versions of the algae-mineral story focuses on whether the constancy of the influence of background causes throughout the study is apparent or not. In the oil-additive story, regardless of whether the scenario was within- or between entities, the influence of background causes on oil runniness was always credibly constant.

Experiment 4.5

Experiment 4.5 began the investigation of causal reasoning on continuous outcomes in a preventive scenario. In this experiment, I replicated Experiment 4.1, except with a scenario

where algae growth before treatment was always higher than after treatment. Note that in any preventive scenario, $PDI_{proportion}$ and $PDI_{multiplication}$ always overlap. Consequently, all experiments involving preventive scenarios technically only investigated two strategies in the judgment task: difference vs. proportion/multiplication (see Table 4.3). Consequently, the scoring analyses only contain two Wilcoxon tests, and the Bonferroni corrected level will be .025.

4.5.1 Method

Participants. 35 undergraduates from School of Psychology, Cardiff University participated to fulfil part of a course requirement. Participants who have participated in Experiment 4.1 were refused from participating in this experiment.

Design and Procedure. This experiment adopted a similar procedure to Experiment 4.1 (within-entity), but with three preventive conditions (see Table 4.3). Also, to reduce any potential confounding between a 'does not influence' judgment and an inability to make an actual judgment, I included the fourth 'Cannot tell' option. If participants opted for this option, they had to elaborate in their own words the reason for their inability to make judgment as described above.

4.5.2 Results

The right side of Table 4.3 shows the results of this experiment. Qualitative inspection indicates that most judgments reflect adoption of a proportion/multiplication strategy; especially in conditions [1.0, 0.5 : 0.6, 0.2] and [0.6, 0.3 : 1.0, 0.5]. In condition [0.8, 0.2 : 0.9, 0.3], the difference based approach appeared to be as popular as the proportion/multiplication strategy. Only two participants consistently used a difference strategy in all three conditions and four participants consistently adopted a proportion/multiplication strategy. Because of this low performance of perfect strategy consistency within participant between all three conditions, I

conducted the scoring analysis on these judgments as explained in Experiment 4.1. Wilcoxon Signed-Ranks Tests of these scores against the chance value of 0.750 show that the proportion/multiplication score (z = -4.490, p < . 001) is significantly above chance, whereas the difference score (z = -.752, p = .452) is not. Vertical-stripe columns on the bottom panel of Figure 4.4 visualize these findings.

4.5.3 Discussion

The scoring analysis supports the qualitative observation that proportion/multiplication was the most prominent strategy in this experiment. In contrast, participants seemed to largely avoid the difference strategy. Because the proportion and multiplication strategy are computationally identical in preventive relations, it was not possible to tease these apart. I continued the investigation by following the rationale of the generative experiments, i.e. by investigating the generalizability of this result in a different context. As discussed in Experiment 4.1 and 4.2, relative to the oil-additive cover story, the algae-mineral cover story was more accommodating towards reasoning about an interaction between background causes and candidate cause. Because of this, generalizability of this result might be an issue between these two stories in preventive scenario. The first step to test this would be the aim of the next experiment, which adapts the oil-additive cover story.

Experiment 4.6

I wanted to test whether perhaps a new cover story could influence the strategy selection, relative to the algae-mineral cover story in Experiment 4.5. More specifically, in this experiment, because the story involved influences of additives on runniness of engine oils, I was interested to assess whether a majority of participants would still use the proportion/multiplication strategy.

4.6.1 Method

Participants. 60 participants were recruited via Amazon Mechanical Turk for USD 0.80. The same participation restriction as in Experiment 4.2 was adopted in this experiment.

Design and Procedure. This experiment used a within-entity situation as in Experiment 4.5 but with a different cover story. This experiment adopted the same cover story as in Experiment 4.2, which was about a study of two additives having influence on runniness of engine oils. A minor modification to the story was that three (as opposed to four in Experiment 4.2) engine oils were involved, corresponding to three conditions. The story highlighted that the study was conducted in a within-entity situation where each oil went through an initial test, followed by mixing with the additives, before undergoing the second test. The rest of the experiment followed the same procedure with Experiment 4.5.

4.6.2 Results

The right side of Table 4.3 summarizes the results. Qualitative observation shows that the proportion/multiplication strategy attracted the most judgments in all conditions. Because perfect strategy consistency across conditions within participants was again low, (with only eight participants consistently adopting the same strategy in all conditions – two participants used difference and six participants used proportion/multiplication), I conducted the scoring analysis as another measure of consistency across the conditions. The scoring analysis corroborated the same trend (see horizontal-stripe column in the bottom panel of Figure 4.4). Wilcoxon Signed-Ranks Tests reveal that the score for proportion/multiplication is significantly greater than chance, $z = -5.435$, $p < .001$, whereas the difference score is not, $z = -1.353$, $p = .176$.

4.6.3 Discussion

The difference strategy was the least adopted among participants in this experiment. Thus, the focus of this discussion is between the proportion and multiplication strategies. From Experiment 4.2, i.e. the generative counterpart of this experiment, I observed that the oil-additive story was less likely to prompt thinking about an interaction between background causes and candidate cause. As evident in that experiment, the proportion strategy was the most adopted, whereas multiplication was the least adopted. Assuming that this story likewise reduced the appeal of causal interactions in a preventive scenario, then leads us to expect that in this experiment, participants would likely reason also with the proportion strategy. Even though I could not conclude whether the algae-mineral or oil-additive context influence reasoning strategy as they did in the generative scenario, the results of both this experiment and Experiment 4.5 converge in showing that the difference strategy was the least relevant in the preventive scenario. In the following two experiments, I extended the investigation to see whether reasoning situation has an influence on strategy adoption as it did in generative scenario.

Experiment 4.7

The objective of this experiment was to study whether the reasoning situation had an influence on strategy selection in preventive scenario of algae-mineral story. In particular, relative to Experiment 4.5, this experiment had a between-entity situation.

4.7.1 Method

Participants. 60 subjects were recruited via Amazon Mechanical Turk for USD 0.80. The exclusion criteria for this experiment were the same as in Experiment 4.2.

Design and Procedure. This experiment was similar to Experiment 4.5 but with modification on the cover story to show a between-entity situation. More specifically, I

mentioned that each laboratory housed two pools: one pool served as a control, and received no treatment, whereas the other pool received treatment with either one (laboratory at location 1) or both (location 2) minerals. A diagram showing this situation (Figure 4.5) further clarified the cover story.

4.7.2 Results

Qualitative observation of the right side of Table 4.3 suggests that most participants again judged according to a proportion/multiplication strategy. This is evident in all conditions except in condition [0.6, 0.3 : 1.0, 0.5], where judgments based on proportion/multiplication were tied with the option based on an unidentified strategy. Similar to Experiment 4.5, perfect strategy consistency across conditions within participant was very low. While only five participants consistently adopted the proportion/multiplication strategy in all conditions, only a single participant used the difference strategy throughout all conditions. The cross-hatched columns in the bottom panel of Figure 4.4 represent the results of the scoring analysis. Wilcoxon Signed-Ranks Tests on the scores show that the proportion/multiplication score ,z = -4.928, p < . 001, is significantly above chance, whereas the difference score is not, z = -1.651, p =.099.

4.7.3 Discussion

Similar to Experiment 4.6, difference was again the least adopted strategy. Hence, the dominant strategy was the proportion or multiplication strategy. Contrasting results of Experiment 4.1 and 4.3 suggested that, in the context of the algae-mineral scenario, a between-entity situation reduced thinking about an interaction between background causes and candidate cause, and that consequently the multiplication strategy was less likely to emerge (see Discussion in Experiment 4.3). Supposing that in the preventive scenario, the between-entity situation also

discouraged interaction-based causal reasoning, I therefore infer that proportion was the most prominent strategy.

Experiment 4.8

This experiment was the parallel to Experiment 4.7 in investigating any possible influence of reasoning situation on strategy selection, but in the oil-additive context. The cover story was the same as in Experiment 4.6 but using a between-entity situation.

4.8.1 Method

Participants. 60 subjects were recruited via Amazon Mechanical Turk for USD 0.80. The exclusion criteria for this experiment were the same as in Experiment 4.2.

Design and Procedure. This experiment adopted the same between-entity as in Experiment 4.7 but using materials of preventive scenario from Experiment 4.6 including the oil-additive cover story and three preventive conditions.

4.8.2 Results

The results of this experiment are on the right side of Table 4.3. Observation of the table reveals proportion/multiplication as the dominant strategy. This is evident in all conditions except in condition [0.6, 0.3 : 1.0, 0.5] where most of the judgments corresponded to a tie between the unknown PDI and proportion/multiplication strategy. Similar to previous experiments, perfect strategy consistency was low with only ten participants regularly using the same strategy. Eight of them consistently adopted a proportion/multiplication strategy, while the remaining two participants applied the difference strategy. The chequered columns at the bottom panel of Figure 4.4 portray the results of the scoring analysis. Wilcoxon Signed-Ranks Tests verify that the proportion/multiplication score is above chance, $z = -5.631$, $p < .001$, whereas the difference score is not, $z = -1.580$, $p = .114$.

4.8.3 Discussion

Experiment 4.6 (the same oil-additive scenario as in this experiment, but using a within-entity situation), Experiment 4.7 (algae-mineral scenario, but using the same between-entity situation as in this experiment) and this experiment converge in revealing difference as the least preferred, and proportion/multiplication as the dominant strategy. Applying the same logic as in the generative counterpart, I interpret these results to indicate that the proportion strategy was the prominent strategy in this experiment as well. This is because, both the oil-additive scenario and the between-entity situation discouraged reasoning about an interaction between background causes and candidate cause, which could lead to adoption of the multiplication strategy. Consequently, even though I cannot in principle disentangle the proportion from the multiplication strategy when considering preventive relations, I interpret this pattern of results to support proportional reasoning as the dominant approach.

4.9 General Discussion of Chapter 4

In this chapter, I presented eight experiments using a paradigm that did not involve asking participants to explicitly compute strength of candidate cause in influencing effect like in previous chapters. The results of these experiments provide further insights on the influence of contexts and reasoning situations on reasoning strategies as evident in Chapter 3.

The four experiments using a generative scenario revealed more complex results relative than their preventive counterparts: Context-specificity (cf. the algae-mineral versus oil-additive scenarios), as well as situation dependency (within- versus between-entity design) were evident. In the algae-mineral context, the reasoning situation moderated strategy use. When in a within-entity situation, the most prominent strategy was multiplication, while in a between-entity situation, the dominant strategy was less apparent and split between difference or proportion.

This moderation, however, was not present in the oil-additive scenario: irrespective of reasoning situation, either within- or between-entity, the majority of participants relied on a proportion strategy.

One possible explanation for variations in these results was the perceived constancy of the influence of background causes on the effect. In other words, it was whether the influence of background causes on the effect remains constant throughout the observation of evidence. A motivation for proposing this factor was due to the inconsistent results between experiments using the oil-additive and the algae-mineral cover stories, especially within the generative scenario. In the oil-additive experiments, it was natural for participants to assume that the influence of background causes on the effect was constant. This is because the story involved a synthetic process that usually took place in a controlled environment. In contrast, participants in the algae-mineral experiments (both generative and preventive scenarios) might think about this differently. Because of the natural, organic setting of the context, it is also natural to think that algae growth happens anyway. This thinking implies that background causes were sufficient in producing the growth, and variably influenced the effect over time: in the beginning, they produced a certain magnitude of the effect, but later their effectivity (power) increased and produced higher magnitude of effect. For example, algae grow naturally in a water body with sufficient exposure to sunlight and plenty of nutrients. As amount of the nutrient increases, the potential for algal bloom also escalate. In short, reasoners in the algae-mineral scenario might have thought that background causes did not influence the effect consistently throughout the observation. When participants thought the influence of background causes remained unchanged throughout the observation, they were more inclined to adopt a proportion strategy. Otherwise,

the results were less straightforward and subject to the influence of reasoning situation (within- or between-entity).

Experiments in preventive scenarios unveiled simpler results. Regardless of context or reasoning situation, these experiments consistently revealed the same result: participants mostly adopted a proportion/multiplication strategy in all experiments. To address this overlapping of proportion and multiplication strategies, I could extend the logic underpinning the influence of scenario and reasoning situation in generative causal inference into preventive scenario, and in general, this extension revealed that the prominent strategy in preventive scenarios would also be proportion-based. This interpretation demands some important precautions, however. This is because of asymmetries between generative and preventive causal inference. Two asymmetries between generative and preventive causal inference that might influence reasoning concern the saliency of an upper limit of causal effectiveness, and differences in the direction of influences between background causes and candidate causes.

Chapter 5: General Discussion

This thesis explores how people reason about causal relations in scenarios where a candidate cause generates or prevents continuous effects. The experimental approach was to conceptually map probabilistic influences in binary causation to (deterministic) influences on continuous effects. This mapping preserved the computational properties of binary causation, which in turn has enabled me to study the applicability of reasoning concepts from binary causation onto causal inference with continuous outcome magnitudes.

In general, results for preventive scenarios were more consistent throughout experiments in all chapters: when reasoning about a candidate cause inhibiting effect magnitudes, reasoners made judgments according to a proportion strategy.[15] In contrast, results for generative scenarios were more complex with evidence of moderating factors.

In Chapter 2, results for the generative scenario indicated that participants reasoned using the difference strategy. Following the suggestion in Buehner et al. (2003), I continued with a series of counterfactual judgment experiments in Chapter 3. While this method was arguably better to elicit reasoning strategies in binary causation, the results of the generative experiments presented here did not support this argument. This is because, despite having manipulations with various contexts and situations, Experiments 3.2 through 3.4 that followed Buehner et al.'s suggestion revealed that most participants simply neglected the base rates and engaged none of the strategies under investigation. The tendency for judgments to reflect a neglecting of base rates, however, was reduced in Experiments 3.5 and 3.6 where, respectively, difference and multiplication were the most prominent strategies. This change in result relative to the former

[15] In Chapter 2 and 3, predictions for multiplication strategy overlapped with proportion.

three experiments could be attributed to the adoption of non-zero counterfactual base rates in these experiments.

Furthermore, contrasting the results between Experiment 3.3 and 3.5, both of which had the same cover story presented in a within-entity setting, indicated that counterfactual base rate (zero vs. non-zero) was a moderating factor of reasoning strategy: i.e. in Experiment 3.3 with zero counterfactual base rate most judgments reflected base rate neglect, whereas in Experiment 3.5 with non-zero counterfactual base rate the results switched to the difference strategy.

In addition, when contrasting the results of Experiments 3.5 and 3.6, both of which adopted non-zero counterfactual base rates, the results were also not consistent. This situation can be attributed to the different context or cover stories, as the next moderator. When a cover story of chemical-algae was used (Experiment 3.5), most participants settled on the difference strategy, whereas in the additive-oil cover story, the leading strategy switched to the multiplication strategy.

Further, in Chapter 4, the experiments highlighted other factors moderating the use of strategy in generative scenarios. As evident from the results of Experiments 4.1 and 4.3, presentation of information in situations of within- versus between-entity moderated reasoning strategy within the same cover story. Moreover, when contrasting the results of Experiment 4.2 relative to 4.1, I observed moderation of context to the adoption of reasoning strategy.

The above results constitute evidence of moderation within the same chapter. When contrasting the results of Experiment 3.6 and 4.2, their results failed to converge (i.e. the multiplication strategy was dominant in the former, while proportion was dominant in the latter) despite both experiments adopting the same additive-oil story in a within-entity format. The most likely explanation for this discrepancy was the method of study: participants made counterfactual

judgments in Experiment 3.6, but implicit judgments in Experiment 4.2. In making counterfactual judgments, participants initially computed a strength index of a cause before applying the index onto a counterfactual situation. In contrast, in the implicit judgments task, participants did not explicitly judge causal strength, but relied on the direction of influence the candidate cause had on effect. This discrepancy evinces that participants responded according to the demands of the tasks.

Perales and Shanks (2008) also showed that demands of the task influence participants' responses accordingly in binary causation: when asked via a counterfactual versus a causal strength question, participants' responses were different. Further, they argued that, when making judgments, people would simply integrate confirming and disconfirming evidence in a linear manner, and proposed a normalized weighted linear combination of the trial-type frequencies, *a*, *b*, *c*, and d (the Evidence Integration, *EI* rule) as the strategy in standard causal learning tasks (for details, see Perales and Shanks, 2007): $EI = \frac{w_a a - w_b b - w_c c + w_d d}{w_a a + w_b b + w_c c + w_d d}$, where *a*, *b*, *c*, d, and w_a, w_b, w_c, w_d correspond to the four cells of a standard 2 x 2 contingency table (see Table 1.1), and their corresponding weightings. Perales and Shanks also argued that when making judgments, depending on how reasoners perceive demands and saliency of information, *inter alia*, they adjust the weights.

Having additional parameters in a model, such as the weightings in the EI rule, is convenient as it offers more degrees of freedom to explain variations in the data, such as the complex results from the generative experiments in this study. Nonetheless, one pertinent issue concerning these additional parameters would be the determination of their values. Further, having additional things to consider would certainly increase cognitive loads. Ultimately, the

question would be whether people actually adopt this approach. In the light of this EI framework, I revisited the results of this study in section 5.2.

In short, participants took into consideration moderating factors when reasoning in generative scenarios, whereas they were consistent throughout all experiments with preventive scenarios. Two possible aspects that might explain the asymmetry of reasoning between generative and preventive scenarios were limit saliency, and direction influence of candidate and background causes.

5.0 Asymmetry of Generative and Preventive

As highlighted in Chapter 1, the limit of the preventive scenario was more salient than in the generative scenario. In preventive situations, because the candidate cause inhibits the effect magnitude, the maximum inhibition is always a complete removal of the effect, i.e. a natural limit of zero (i.e. in the oil-mineral story, this would mean that the mineral thickens the oil, to the extent that it does not drip at all). Such a salient limit is not always present in a generative scenario, as the candidate cause can keep generating the effect magnitude as long as it has the opportunity to do so. While this low saliency of limit might have been an issue for the reasoning process especially in the generative scenario, nonetheless, the fact that a large number of judgments corresponded to the proportion strategy indicated that participants were aware of the artificial limit I deployed, and used it accordingly. For comparison, in binary probabilistic causation, the saliency of the limit is not an issue: the relevant information has clear limits of zero and one, marking the maximum effectiveness for preventive and generative scenarios respectively.

Another asymmetry between generative and preventive scenario structures concerns how background causes and candidate cause influence the effect. In generative scenarios, these two

causes influenced the effect in the same direction (they both generate algae growth/runniness of oil), whereas in the preventive scenarios, their influences on the effect opposed each other (background causes generate growth/runniness, but candidate causes inhibit it). When the background causes influence the effect in the *same* direction as the candidate cause (as is the case in generative scenarios), it might be less appealing to consider background and candidate cause as acting independently of each other. In other words, candidate and background might be more likely to be seen as relating to and influencing each other in a generative scenario. In contrast, in the scenario where they oppose each other (i.e. preventive), background causes and candidate cause may a priori appear to be independent of each other.

In short, these two asymmetries between generative and preventive scenarios differentially impact the three reasoning strategies under study: The proportion strategy relies on a consideration of an upper limit – which does not naturally exist for continuous outcome magnitudes in generative scenarios – and the difference and multiplication strategies do not; in preventive scenarios, the upper limit of causal effectiveness is automatically implied to be a reduction to zero magnitude, and thus the proportion strategy is on equal footing with the two other strategies in this respect.

Secondly, while the proportion strategy demands an assumption of independence between candidate and background causes, the other two strategies do not. In fact, multiplication-based reasoning by definition is based on thinking of an interaction between candidate and background causes. As I discussed above, preventive scenarios may lend themselves more naturally to the assumption of candidate and background acting independently from each other than generative scenarios, so again the proportional strategy might be more readily accessible in preventive than in generative scenarios.

5.1 Multiplication Strategy in Binary Causation

Having a less salient limit, and a weaker sense of independence between influences of background causes and candidate cause on the effect, generative causal relations with continuous magnitude may be more susceptible to reasoning via causal interaction, which is the basis of the multiplication strategy. The multiplication strategy, which I found in the generative scenario during my pilot study, is not discussed in any literature that I am aware of. One possible explanation for the absence of multiplication strategy in binary causation lies in its reasoning structure. Being probabilistic, reasoning with binary causal relations emphasizes judgments from a *group of instances*. Thus, what matters in probabilistic judgment is the *number of instances* where the effect is present, while the *extent* of the effect is not the focus and is normally not being considered (but see Lovibond, Been, Mitchell, Bouton & Frohardt, 2003).

In contrast, the work presented in this study considered judgments about *an individual instance*, and put focus on the *extent* of the influence occurring on that instance. Because an interaction between background causes and candidate cause to influence the effect involved changes on the effect magnitude (i.e. the extent of the effect), the interaction is more noticeable at the individual instance level instead of at the group level. Since the multiplication strategy stems from this causal interaction, therefore, this strategy surfaces only at the individual level. Consequently, binary causation, which focuses only at the group level, does not pick up the multiplication strategy.

5.2 Revisiting Results: In the Light of Evidence Integration Rule

To revisit the results, firstly, I need to consider how to apply the probabilistic binary framework of EI rule into this continuous framework. The main idea of EI rule in binary causation focuses on contrast of confirmatory against disconfirmatory information in 2 x 2

contingency table: i.e. information in cells *a* and *d* are confirmatory, while *b* and *c* are not (for details see Perales and Shanks, 2007). Because the current study does not consider information from individual cells, but only a direct mapping of $\frac{a}{a+b}$ (i.e. P(e|c)) into Q(e|c) and $\frac{c}{c+d}$ (i.e. P(e|c)) into Q(e|¬c), I could not exactly revisit the results with respect to EI.

Mathematically, the main idea of EI rule is to consider the difference between confirmatory and disconfirmatory evidence, relative to the total evidence observed. Thus, with respect to my work, I could consider the effect magnitude when the cause is present, i.e. Q(e|c), as confirmatory, and the magnitude when cause is absent, i.e. Q(e|¬c), as disconfirmatory; and apply the concept. In other words, the adjusted EI rule to suit the current study can be summarised in this equation, $\mathrm{EI} = \frac{\mathrm{Q(e|c)} - \mathrm{Q(e|\neg c)}}{\mathrm{Q(e|c)} + \mathrm{Q(e|\neg c)}}$. While this does not exactly map the binary EI rule to the continuous magnitudes of this study, this proposal partially holds some merit because this is the only information that was available to the participants during the experiments. While the original version of the model has factored in weightings of each individual cells of contingency table, I only considered the un-weighted version, as discussed above, when revisiting the results of tendency analysis for all experiments in Chapter 3.

Revisiting Chapter 3. The revisited results from experiments in this chapter are in the right-most column of Table 5.1. From all experiments in Chapter 3, the EI adapted strategy would make a difference to only Experiment 3.8 (highlighted cell of Table 5.1): In this experiment, the proportion/multiplication and EI adapted strategies scored the same in the tendency analysis. This suggests that participants in Experiment 3.8 were equally likely to adopt either of these two strategies. In other experiments, the EI adapted strategy mostly scored the lowest.

Table 5.1: Revisited Table 3.1 with results of adapted EI rule on the data

Generative Exp.	Cover story		Means (*SD*) for Strategies				
	Theme	Design	Base rate Neglect	Proportion Strategy	Difference Strategy	Multiplication Strategy	EI rule adaptation
3.1	Cream-Rash	W	.24(.02)	.40(03)	.69(.05)*	-	.35(.03)
3.2	Fertilizer-Crop	B	.60(.05)*	.45(.03)	.40(.04)	-	.15(.03)
3.3	Chemical-Algae	W	.53(.03)*	.42(.02)	.46(.03)	-	.17(.02)
3.4	Chemical-Algae	B	.62(.09)*	.47(.05)	.44(.08)	-	.18(.04)
3.5	Chemical-Algae	W	.37(.03)	.39(.03)	.54(.04)*	.50(.04)	.13(.01)
3.6	Additive-Oil	W	.13(.01)	.38(.02)	.37(.02)	.59(.04)*	.14(.01)
Preventive Exp.			Base rate Neglect	Proportion/ Multiplication Strategy	Difference Strategy		EI rule adaptation
3.7	Cream-Rash	W	.31(.02)	.52(.04)*	.41(.02)		.16(.02)
3.8	Chemical-Algae	W	.57(.03)	.58(.03)*	.42(.02)		**.58(.03)***
3.9	Chemical-Algae	B	.67(.08)*	.50(.06)	.33(.04)		.50(.06)
3.10	Additive-Oil	W	.31(.02)	.52(.02)*	.41(.02)		.17(.01)

Note: 'W' and 'B' in the third column refer 'within-entity' and 'between-entity' design as explained in the text. The higher the score, the more inclined participants were towards that strategy. The means in this table refer to across participants' average scores in each experiment.

Revisiting Chapter 4. Revisited results for all experiments in this chapter were more interesting. This is because the predictions of the EI adapted strategy for all conditions used in the experiments completely overlap with predictions of the multiplication strategy. Table 4.2 and 4.3 shows the predictions and results of all strategies for all conditions in this chapter. Thus, considering only results of this chapter, I could infer that participants were likely to adopt multiplication and EI adapted strategy. While this provides a merit for the EI adapted strategy, a more conclusive argument would need a proper experiment because of the conditions: The conditions used in these experiments were tailored such that their directions of predictions do not completely overlap with each other, whenever possible.

Considering both revisited results from Chapter 3 and 4, there are some traces of support for the EI adapted strategy. I could not, however, make any conclusion on its applicability to explain causal reasoning with continuous outcomes in this study. This is because the adaptation from the original probabilistic version was simplified, as discussed at the beginning of 5.2 sub-chapter. Further, the un-weighted version of the model that I used assumed both confirmatory and disconfirmatory evidences are equally salient, which White (2003) argues to be unlikely. Perhaps a weighted model would produce a better fit, but doing so would also required me to adjust weightings for the proportion and difference strategies as well. In short, the un-weighted version of EI rule provides a comparison with other strategies at the same level.

5.3 Summary and Conclusion

The work I presented in this thesis built upon a rich research tradition into human reasoning with binary causal relations, and extended it by directly mapping extant paradigms for the study of reasoning with binary probabilities on to continuous deterministic relations. Despite being restricted to deterministic settings, the contribution of this work is to address whether reasoning about relations with continuous outcomes is rooted in the same processes as reasoning about binary relations. The absence of a consistent strategy within participants and within conditions across many experiments would suggest that neither a Power PC-rooted proportion strategy, nor a ΔP-rooted difference strategy successfully explains reasoning about causal relation with continuous outcomes.

5.4 Future Direction

The current work did not fully explore continuous causation (i.e. relationships between two continuous variables). A simple continuous relation is three-dimensional: a continuous cause, a continuous effect, and the probability of the effect (see Young and Cole, 2012). While

the work of Young and Cole gives an indication that people are able to deal with this degree of complexity, their work did not attempt to identify any framework, however.

Thus, the need to consider a new theory or framework is apparent. Consider this scenario: Imagine there are two pools (pool A with algae growth of 10 m^2, and pool B with algae growth of 20 m^2) and two different chemicals of interest. After administering a same amount of chemical A and chemical B (e.g. 5 litres) into the respective pool, algae growth in pool A increases its coverage to 60 m^2, while in pool B it increases to 80m^2. If I consider which of these two chemicals has higher effectivity in producing algae growth, a natural response would be to compute causal strength indices of chemical A and B, and compare them. Because the amount of chemicals is the same in both cases, solving these causal relations would be relatively easy. In this case, a reasoner would intuitively consider the difference of growth before, and after the administration to conclude that chemical B (80 – 20 = 60) is stronger over A (60 – 10 = 50).

If the scenario were a bit different, where the amount of chemicals administered were not the same, e.g. 5 litres of chemical A, and 10 litres of chemical B, the judgment would be harder. This is because reasoners need not to only consider the degree of change of effect magnitude before and after the chemical administration, but also the degree of change of the chemicals administered into the pools. My proposal for this is to consider the degree of change for the effect with respect to degree of change for the cause. In this case, 50m^2 change of algae in pool A is attributed to a 5 litre change of chemical, whereas in pool B 60m^2 change of algae is attributed to a 10 litre change of chemical. A possible approach to represent the strength of chemical would be to consider to what extent each chemical could produce algae if only 1 litre had been administered: computationally, it is a proportion of growth relative to amount. For chemical A this would be 10m^2/l (i.e. $\frac{50}{5}$), and 6m^2/l (i.e. $\frac{60}{10}$) for chemical B.

Mathematically, the above proposal is a computation of slope for a linear relation. Figure 5.1 visualises this. In the figure, the x-axis (horizontal) represents amount of the chemical (litre), while the y-axis (vertical) represents algae growth (meter^2). Further, lines A and B represent the relations between chemicals A and B in their respective pools. Visually, the slope for line A is steeper than line B, which is consistent with the earlier computation that strength indices for chemical A and B are respectively 10m^2/l and 6m^2/l.

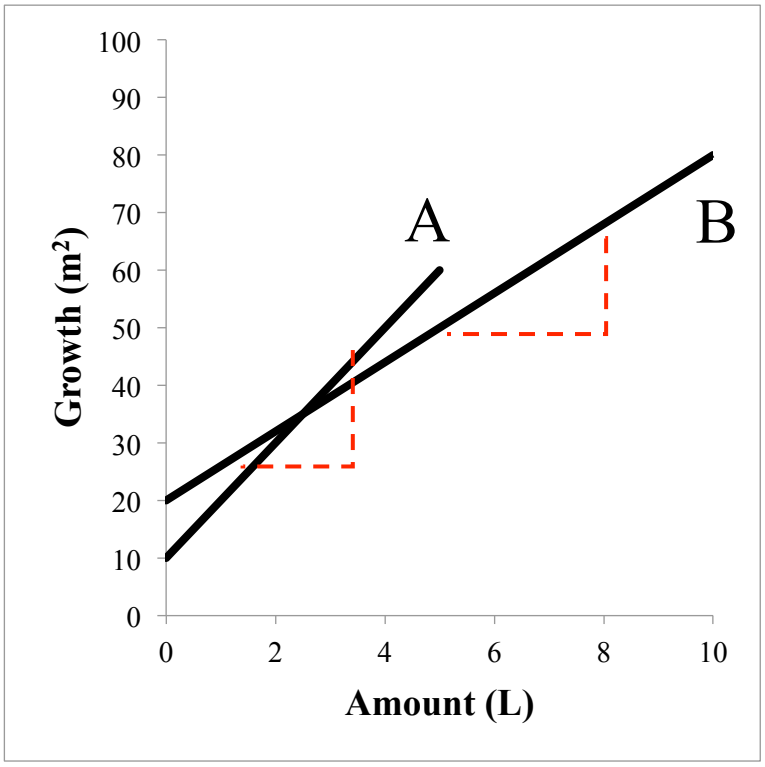


Figure 5.1. Visualisation of scenario 7 example

While this proposal seems intuitive, it is premature because of two issues: it addresses only non-probabilistic computation, and linear relations. To consider the probability of effect, the graph in Figure 5.1 would need to have the third z-axis (depth). This new dimension contains probability information of every point on both line A and B. The second issue with this proposal is that it captures only linear relations. While people often assume any relation as linear and

positive (Brehmer, 1992, 1994; Carroll, 1963; Diehl & Sterman, 1995), there could be other relations such as quadratic, logarithmic or power. The literature on function learning has highlighted that it is easier for people to learn linear functions than quadratic (Carroll, 1963), power than linear or logarithmic (Koh & Meyer, 1991), and noncyclic than cyclic ones (Byun, 1995), *inter alia*. These works suggest that despite the complexity of how two continuous variables are related, people have capability to identifying the function.

With all of these complexities, a natural question would be: what is the way forward? In the next paragraphs, I would like to answer this question. For simplicity, lets consider only linear relations between a continuous cause and a continuous effect. Given the support Power PC theory has received in the literature on binary causation, I think its core concept of causal reasoning is promising. Thus, my proposal of the new theory in continuous causation is in line with this theory.

When proposing the Power PC theory, Cheng (1997) highlighted that its formation involved an interpretation of a reasoner who infers the magnitude of the unobservable causal power from observable events based on his or her theoretical explanation as well as belief about alternative causes. Specifically, she considered the probability of unobservable power of *i* to coincide with the observable probability of effect *e* in the presence of candidate cause *i*, when no other cause is present or exists[16]. Analogously, if I were to follow the same idea, perhaps a framework for continuous causation between a linear relation of *i* and *e* is to consider the probability of unobservable power of *i* exerted on effect *e* to coincide with the probability that the relation has a particular slope of *m*, when other causes are absent. The probability of *m* is a

[16] Some researchers use causal power to refer to causal mechanism, or propensity from a source to produce the effect. Cheng (1997) defines causal power as "*the intuitive notion that one thing causes another by virtue of the power or energy that it exerts over the other*".

measure of stability of candidate cause *i* to consistently produce effect *e* according to a pattern (in this case linear function).

This line of thought, however, assumes that reasoners had to learn how the candidate cause relates with the effect beforehand. The argument that Griffiths and Tenanbaum (2005) brought forth in binary causation could support this: prior to addressing the *strength* with which a candidate cause influences an effect, people first need to identify whether there exists a causal relation between the cause and effect, i.e. the *structure*. Hence, in continuous causation, to identify whether a relation between two continuous variables exists would mean to identify their functional form, i.e. how the two variables are related. In other words, the parallel idea for Griffiths and Tenanbaum's causal structure in binary causation would be the function learning in continuous causation.

In this framework proposal, two crucial properties that may influence reasoning strategy of continuous causation are the relational index (in the case of linear relation, I suggest the slope as the index), and the consistency of the index across many instances (i.e. its probability). Of course, the more important question is how these two properties interact: The higher the index, the stronger *i* and *e* are related; the higher the probability of the index the more stable that relation is.

The main focus of this thesis was on causal relations involving continuous outcomes, instead of fully exploring continuous causation (i.e. relation between two continuous variables). Therefore, even though the results indicated that neither the proportional nor difference strategy were capable to capture reasoning strategy in this kind of causal relation, these two strategies may still be relevant when considering the full scope of continuous causation.

References

Allan L. G. (1980). A note on measurement of contingency between two binary variables in judgement tasks. *Bulletin of Psychonomic Society*, *15*, 147–149.

Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition*, *23*, 510–524.

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*, 1119 - 1140.

Cheng, P. W. (1993). Separating causal laws from causal facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 30, pp. 215-264). San Diego: Academic Press.

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

Cheng, P. W. & Buehner, M. J. (2012). Causal Learning. In Holyoak, K. J. & Morrison, R. G. (Eds.) The Oxford Handbook of Thinking and Reasoning. Oxford University Press, Oxford, England, pp. 210 – 233.

Collins, D. J., & Shanks, D. R. (2006). Conformity to the Power PC theory of causal induction depends on the type of probe question. *Quarterly Journal of Experimental Psychology, 59,* 225–232.

Gigerenzer, G., & Todd, P. M. (1999). Simple heuristics that make us smart. New York: Oxford University Press.

Griffiths, T. L., Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Hume, D. (1987). *A treatise of human nature (2nd ed.)*. Oxford, England: Clarendon Press. (Original work published 1739).

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and applied*, 79(1, Whole No. 594).

Kant, I. (1965). Critique ofpure reason. London: Macmillan. (Original work published 1781)

Liljeholm, M., & Cheng, P. W. (2007). When Is a Cause the ‘‘Same’’? Coherent Generalization Across Contexts. *Psychological Science, 18*, 1014 - 1021.

Lovibond, P. F., Been, S., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition*, *31*, 133–142.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*, 249-26.

Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge, UK: Cambridge University Press.

Perales, J. C., & Shanks, D. R. (2003). Normative and descriptive accounts of the influence ofpower and contingency on causal judgement. *Quarterly Journal of Experimental Psychology, 56*, 977 - 1007.

Perales, J. C., & Catena, A. (2006): Human causal induction: A glimpse at the whole picture, *European Journal of Cognitive Psychology*, *18*, 277-320.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: a review and synthesis. *Psychonomic Bulletin & Review, 14*, 577 - 596.

Perales, J. C., & Shanks, D. R. (2008). Driven by Power? Probe Question and Presentation Format Effects on Causal Judgment. *Journal of Experimental Psychology: Association Learning, Memory, and Cognition*, *34*, 1482–1494.

Sloman, S. A. (2005). Causal models: How people think about the world and its alternatives. Oxford: Oxford University Press.

Sloman, S., & Lagnado, D. (2005). Do we "do"? *Cognitive Science*, *29*, 5 - 39.

Tolman, E. C., & Brunswik, E. (1935). The organism and the causal texture of the environment. *Psychological Review*, *42*, 43-77.

Ward W. C., Jenkins H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology*, *19*, 231–241.

Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the encironment. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 26, pp. 27–82). San Diego, CA: Academic Press.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response- outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 174 - 188.

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

White, P. A. (2003). Making causal judgments from the proportion of confirming instances: The pCI rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 710-727.

Wu, M. & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, *10*, 92 - 97.

Young, M.E., & Cole, J.J. (2012). Human sensitivity to the magnitude and probability of a continuous causal relation in a video game. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 11-22.

Appendix A: Cover Stories for All Experiments

A.1 Cover Stories For Experiment 2.1

A.1.1 For Generative-Continuous Experiment in Chapter 2 and Experiment 3.2

Imagine that you are an agricultural consultant conducting a study on the effectiveness of different fertilizers on the growth of corn. In the European Union, the corn crops planting season begins in mid-April, and harvesting runs from mid-August through late October.

Corn requires a particular composition of nitrogen, phosphate and potash to produce good yield. An imbalance of these nutrients results in soils that have an unsuitable chemical make up that may hinder the growth of the corn plants. Therefore, fertilizers are administered to balance the nutrient composition in the soil, both to optimize the growth of the plant, and also to prevent dying off of young corn seedlings. Because different fertilizers have different nutrient compositions, fertilizers may differ in the extent to which they boost corn yield. Yields can be estimated by measuring the total area in which usable corn crops can be harvested with respect to the total planted area.

In this study you are assessing the effectiveness of 15 different fertilizers to improve corn yield. The study was an EU-wide effort, where 15 different EU member states each tested the effectiveness of ONE particular fertilizer. In each EU country a particular fertilizer was administered to an experimental plot of 100 meters square, freshly sown with corn. Each country also employed a control plot, also of 100 meters square, where corn was sown and grown in the absence of any fertilizers. The total area of usable corn crops in each field has been monitored and recorded throughout the study period.

You will see the records from the 15 participating countries, each testing one of the 15 different fertilizers. For each dataset, you have to judge the strength of the effectiveness of each

fertilizer in maximizing corn yield by selecting a number on the strength scale. The scale begins with 0 referring to “absolutely ineffective” and ends at 10 referring to “absolutely effective”.§

A.1.2 For Generative-Binary Experiment

Imagine that you are an agricultural consultant conducting a study on the effectiveness of different fertilizers on the growth of corn. In the European Union, the corn crops planting season begins in mid-April, and harvesting runs from mid-August through late October.

Corn requires a particular composition of nitrogen, phosphate and potash to produce good yield. An imbalance of these nutrients results in soils that have an unsuitable chemical make up that may hinder the growth of the corn plants. Therefore, fertilizers are administered to balance the nutrient composition in the soil, both to optimize the growth of the plant, and also to prevent dying off of young corn seedlings. Because different fertilizers have different nutrient compositions, fertilizers may differ in the extent to which they boost corn yield. Yields can be estimated by counting the number of corn plots in which usable corn crops can be harvested with respect to the total planted plots.

In this study you are assessing the effectiveness of 15 different fertilizers to improve corn yield. The study was an EU-wide effort, where 15 different EU member states each tested the effectiveness of ONE particular fertilizer. In each EU country, a particular fertilizer was administered to 10 experimental plots of freshly sown corn. Each county also employed 10 control plots where corn was sown and grown in the absence of any fertilizers. You will receive information about experimental and control plots in each country, telling you whether or not a particular plot showed improved yield.

You will see the records from the 15 participating countries, each testing one of the 15 different fertilizers. For each dataset, you have to judge the strength of the effectiveness of each

fertilizer in maximizing corn yield by selecting a number on the strength scale. The scale begins with 0 referring to “absolutely ineffective” and ends at 10 referring to “absolutely effective”.

A.1.3 For Preventive-Continuous Experiment

Imagine that you are an oncologist conducting research on the effectiveness of new chemical agents in fighting tumours. A tumour is an abnormal mass of tissue as a result of abnormal proliferation of cells that form a lump on a particular location of the body. This abnormal cell proliferation is the result of genetic mutations that occur when the genome cells are exposed to radiation, viruses, transposons and mutagenic chemicals. While certain types of tumours have a relatively low medical risk, other types do impose high fatality risk.

Chemotherapy is one of the methods to fight tumours by administering chemotherapeutic agents. Chemotherapeutic agents work by impairing the cell division process and effectively preventing the growth of the tumour. Once a tumour has been diagnosed, it can grow up to a certain volume. Administering the chemotherapeutic agent can reduce the tumour size by preventing its growth. Some chemotherapeutic agents are more effective in preventing tumour growth than others. Thus, the purpose of this study is to investigate the effectiveness of various chemotherapeutic agents in preventing the tumour growth.

In this study, there were 15 different labs participating, and each studied ONE particular chemotherapeutic agent. Every lab was given two groups of mice that were exposed to radiation. Previous studies on radiation-induced tumour growth suggest that - in the absence of any preventive treatment - mice should develop a brain tumour of 10 micrometres cubic within one week.

In each study, a control group of mice was exposed to the radiation treatment and the tumour size was measured after one week. An experimental group of mice received exactly the

same radiation treatment, but was administered the chemotherapeutic agent during the week, and tumour size was measured at the end of this week as above.

Your task is to go through records of average tumour sizes from the 15 labs, each investigating one of the 15 types of chemotherapeutic agents. For each record, you have to judge the strength of its effectiveness in preventing the growth of the tumours by selecting a number on the strength scale between 0, referring to "absolutely ineffective" and 10, referring to "absolutely effective".

A.1.4 For Preventive-Binary Experiment

Imagine that you are an oncologist conducting research on the effectiveness of new chemical agents in fighting tumours. A tumour is an abnormal mass of tissue as a result of abnormal proliferation of cells that form a lump on a particular location of the body. This abnormal cell proliferation is the result of genetic mutations that occur when the genome cells are exposed to radiation, viruses, transposons and mutagenic chemicals. While certain types of tumours have a relatively low medical risk, other types do impose high fatality risk.

Chemotherapy is one of the methods to fight tumours by administering chemotherapeutic agents. Chemotherapeutic agents work by impairing the cell division process and effectively preventing the growth of the tumour. Some chemotherapeutic agents are more effective in preventing tumour growth than others. Thus, the purpose of this study is to investigate the effectiveness of various chemotherapeutic agents in preventing the tumour growth.

In this study, there were 15 different labs participating, and each studied ONE particular chemotherapeutic agent. Every lab was given two groups of twenty mice that were exposed to radiation. Previous studies on radiation-induced tumour growth suggest that - in the absence of any preventive treatment - mice should develop a brain tumour within one week.

In each study, a control group of mice was exposed to the radiation treatment and the number of mice with tumour was recorded after one week. An experimental group of mice received exactly the same radiation treatment, but was administered the chemotherapeutic agent during the week, and number of mice with tumour was recorded at the end of this week as above.

Your task is to go through records from the 15 labs, each investigating one of the 15 types of chemotherapeutic agents. For each record, you have to judge the strength of its effectiveness in preventing the growth of the tumours by selecting a number on the strength scale between 0, referring to "absolutely ineffective" and 10, referring to "absolutely effective".

A.2 Cover Stories For Experiment 2.2

A.2.1 For Clear-Limit Experiment

The opening ceremony of the 2008 Summer Olympic was not just a great event, but a historical day for the Beijing Weather Modification Office. The Beijing Weather Modification Office reported that "shooting" down all possible clouds heading to the city resulted in the ceremony passing by without a single drop of rain and with temperatures never exceeding 29 degrees Celsius. In the United States a similar technology has been studied over a 12 000 square kilometres, with the goal to influence the amount of rainfall, length of storms, and the area in which rain falls.

In cloud seeding technology, normally silver iodide is sprayed into the cloud to initiate a continuous merging process of tiny water drops, resulting in bigger and heavier drops that fall down as rainfall. However, silver iodide is suspected to lead to health related side effects, which has become a major concern. Consequently, safer alternatives are being researched

Imagine that you are a meteorologist investigating new chemical agents that might substitute silver iodide in the cloud seeding procedure. You conducted a study to determine the

effectiveness of various chemical agents to produce rainfall. Your investigation took place in 15 different countries, each studying ONE chemical agent. In each country, a cumulus cloud at an experimental location were sprayed with the chemical agent, whereas another cloud at a geographically similar and nearby control location were left untouched.

In each study, the relative humidity inside each of the cloud areas was measured. The maximum possible relative humidity is 100%, which means that humidity is so high that water will condensate and fall down as rain.

Your task is to review the 15 records of the humidity levels in both locations, from every country. For each record, you have to judge the strength of the chemical in maximizing relative humidity by selecting a number on the strength scale between 0, referring to "absolutely ineffective" and 10, referring to "absolutely effective".

A.2.2 For No-Limit Experiment

The opening ceremony of the 2008 Summer Olympic was not just a great event, but a historical day for the Beijing Weather Modification Office. The Beijing Weather Modification Office reported that "shooting" down all possible clouds heading to the city resulted in the ceremony passing by without a single drop of rain and with temperatures never exceeding 29 degrees Celsius. In the United States a similar technology has been studied over a 12 000 square kilometres, with the goal to influence the amount of rainfall, length of storms, and the area in which rain falls.

In cloud seeding technology, normally silver iodide is sprayed into the cloud to initiate a continuous merging process of tiny water drops, resulting in bigger and heavier drops that fall down as rainfall. However, silver iodide is suspected to lead to health related side effects, which has become a major concern. Consequently, safer alternatives are being researched

Imagine that you are a meteorologist investigating new chemical agents that might substitute silver iodide in the cloud seeding procedure. You conducted a study to determine the effectiveness of various chemical agents to produce rainfall. Your investigation took place in 15 different countries, each studying ONE chemical agent. In each country, a cumulus cloud at an experimental location were sprayed with the chemical agent, whereas another cloud at a geographically similar and nearby control location were left untouched. The amount of rainfall from clouds at both locations were measured (in millimetres) and recorded.

Your task is to review the 15 records of the clouds in both locations, from every country. For each record, you have to judge the strength of the chemical agent in maximizing the amount of rainfall by selecting a number on the strength scale between 0, referring to "absolutely ineffective" and 10, referring to "absolutely effective".

A.2.3 For Binary Experiment

The opening ceremony of the 2008 Summer Olympic was not just a great event, but a historical day for the Beijing Weather Modification Office. The Beijing Weather Modification Office reported that "shooting" down all possible clouds heading to the city resulted in the ceremony passing by without a single drop of rain and with temperatures never exceeding 29 degrees Celsius. In the United States a similar technology has been studied over a 12 000 square kilometres, with the goal to influence the amount of rainfall, length of storms, and the area in which rain falls.

In cloud seeding technology, normally silver iodide is sprayed into the cloud to initiate a continuous merging process of tiny water drops, resulting in bigger and heavier drops that fall down as rainfall. However, silver iodide is suspected to lead to health related side effects, which has become a major concern. Consequently, safer alternatives are being researched

Imagine that you are a meteorologist investigating new chemical agents that might substitute silver iodide in the cloud seeding procedure. You conducted a study to determine the effectiveness of various chemical agents to produce rainfall. Your investigation took place in 15 different countries, each studying ONE chemical agent. In each country, 20 cumulus clouds at an experimental location were sprayed with the chemical agent, whereas another 20 clouds at a geographically similar and nearby control location were left untouched. Observations of whether rain fell in these two locations were recorded.

Your task is to review the 15 records of 20 clouds in both locations, from every country. For each record, you have to judge the strength of the chemical agent in producing the rainfall by selecting a number on the strength scale between 0, referring to “absolutely ineffective” and 10, referring to “absolutely effective”.

A.3 Cover Stories For Experiment 3.1

Imagine that you are a pharmaceutical consultant researching the side effects of synthetic substances in cosmetic creams. One common side effect in cosmetic products is irritant contact dermatitis. The symptom usually appears as skin rash but can develop to blisters if left untreated. Skin rash is a change to the skin causing it to become reddish, bumpy, itchy and sometimes painful. During the irritant reaction, the immune system fights back, as if the cosmetic cream is harmful, by producing histamine that is responsible for the skin rash.

Skin rash normally occurs at the area of contact with the cream. Areas where the outermost layer of skin is thin, or dry and cracked are more susceptible to skin rash. Your task is to investigate how strongly various cosmetic creams cause a skin rash as a side effect.

In this study, there were 15 different labs participating, and each studied ONE particular cosmetic cream. Naturally, some people may develop skin rash even in the absence of any

cosmetic products. Thus, your study follows a before-after design where in each lab a participant's skin is examined for rash before the cream is applied, and then checked again afterwards.

In each lab, the patient's back was examined for rash, and the size of any rash present was recorded. Then the cosmetic cream was applied on the back of the patient to cover 10 centimetres squares of the skin area. After one hour, the area of skin rash, where the cream was applied, was measured and recorded.

You will see the records from the participating 15 labs, each investigating one of the 15 types of cosmetic creams. For each cream, we are asking you to consider how strongly you think it causes skin rash as a side effect. To do so, we are asking you to imagine a new patient who does not suffer from a rash. We are then asking you to imagine how much the skin rash area would be, once the cream would be applied to a certain area on the patient's back.

A.4 Cover Stories For Experiment 3.3 and 3.8

Carbon dioxide is an important component in the Earth's atmosphere because it plays a significant role in the greenhouse effect. One way to reduce carbon dioxide is by sequestering it underground where it can mineralize and turn into solid carbonate minerals. Unfortunately, this is a very slow process. Nonetheless, scientists have recently discovered the use of bacteria to expedite this sequestering process.

Imagine that you are a biochemist investigating the influence of chemicals on microbiological ecosystems. You have found a group of chemicals that can effectively increase the amount of carbon-sequestering bacteria in an aquatic system. Although these chemicals have already been shown to increase carbon-sequestering bacteria populations, they may have a side effect that can influence algae populations.

Algae are a very common group of micro-plants that can be found in most aquatic systems. Algae growth rate depends on the environmental factors, such as temperature, inorganic chemical nutrients, and salinity. Rapid algae growth can result in a phenomenon known as algal bloom, in which algae concentrations may reach millions of cells per milliliter. In this state, the production of phytoplankton, as the base form of the marine food chain, is disrupted, hence endangering the whole ecosystem.

You were hired by a company to study the side effects of the chemicals with relation to algal bloom. Specifically, you were looking at the relationship of the chemicals in *increasing* [*preventing* in Experiment 3.8] algae growth, as a side effect, in a fixed amount of time.

The company wanted to study algae growth in natural water reservoirs, so they identified 15 different lakes suitable for the study. One lab was built nearby each lake beforehand. The company had requested that each lab study only ONE particular chemical substance and its effect on algae growth. An indoor pool was set-up in each lab to control for other potential variables, and water from the lake was drawn to fill the pools. The sizes of the pools and volumes of water from the lakes were identical across all pools and labs. Each pool in the study was built to have a surface area of 100 square meters.

You have visited each lab and checked whether and how much of the pool surface area was covered with algae. The chemical was then administered via aerial spray over the pool. Naturally, algae grow at different rates, depending on various conditions. This means that most likely the amount of algae growth will vary across different labs. Critically, however, the conditions WITHIN each lab were kept constant, so that such variation would affect the pool to the same extent before and after the treatment period. After two weeks, the surface area of the pool that was covered with algae was re-measured.

Your task is to go through records from the 15 labs. Each record contains information on the surface area covered by algae in the pool before and after treatment with the chemical in question. For each record, we are asking you to consider how strongly the chemical *causes* [*prevents* in Experiment 3.8] algae growth as a side effect.

To do so, we are asking you to imagine another lab that also has a pool inside. We are then asking you to imagine how much of the area of the pool would be covered with algae, once the chemical substance would be administered.

A.5 Cover Stories For Experiment 3.4

Algae are a very common group of microplants that can be found in most aquatic systems. Algae, like other plants are autotrophs: they are capable of converting sunlight into food to stay alive. Algae can reproduce in either sexual or asexual, or both ways, depending on the environmental conditions they are in. Environmental factors, such as temperature, inorganic chemical nutrients, and salinity, regulate not only the method, but also the rate of reproduction.

Algae's characteristics of ubiquitousness and high content of lipid (i.e. oil) have received significant attention from biofuel enthusiasts. Lipid can be harvested from algae's cell walls and can be used as biofuel to power a wide range of machines. Consequently, there is now a research drive on how to grow algae en masse. The best way to cultivate algae for this purpose is in indoor settings, which prevent damage from harsh weather conditions and cross-fertilization with other, less productive strands.

Imagine that you are a biochemist investigating the influence of chemical substances on algae reproduction. Your investigation strives to prepare the most suitable chemical environment for algae to grow and mature in a fixed amount of time.

In this project, 15 different labs have participated, and each lab studied ONE particular chemical substance. Two indoor pools were set-up in each lab to control for other potential variables. The sizes of the pools and volumes of water filled in were identical in all pools and labs. Each pool in the study was built to have a surface area of 100 square meters.

In each lab, both pools were seeded with algae cysts. The first pool was retained as a control, whereas the chemical substance was then administered via aerial spray over the second pool. The two pools were in separate sections of each lab to ensure that the treatment could not affect the control pool. Other than the treatment though, the conditions in both pools were identical (i.e. chemical, biological, and physical characteristics of water and surrounding air, temperature, etc.). Naturally, algae mature and reproduce at different rates, depending on various conditions. This means that most likely the amount of algae growth will vary across different labs. Critically, however, the conditions WITHIN each lab were kept constant, so that such variation would affect both the control and treatment pools to the same extent. After two weeks, the area of water surface covered with matured algae in each pool was measured.

Your task is to go through records from the 15 participating labs. Each record contains information on the surface area covered by matured algae in the pools treated with the chemical in question, and the control pool without treatment. For each record, we are asking you to consider how strongly the chemical causes algae reproduction.

To do so, we are asking you to imagine a new pool seeded with algae cysts. We are then asking you to imagine how much of the area of the pool would be covered with matured algae, once the chemical substance would be administered.

A.6 Cover Stories For Experiment 3.5

Imagine that you are a biochemist investigating the influence of chemicals on microbiological ecosystems. You have found a group of phosphoric salts that can effectively increase the amount of carbon-sequestering bacteria in an aquatic system. Although phosphoric salts have already been shown to increase carbon-sequestering bacteria populations, they may have a side effect that can influence algae populations.

Algae are a very common group of microplants that can be found in most aquatic systems. Algae growth rate depends on various environmental factors. Rapid algae growth can result in a phenomenon known as algal bloom, in which algae concentrations may reach millions of cells per milliliter. In this state, the production of phytoplankton, as the base form of the marine food chain, is disrupted, hence endangering the whole ecosystem.

You were hired by a company to study the side effects of 15 different varieties of phosphoric salts with relation to algal bloom. Specifically, you were looking at the relationship of the chemicals in causing algae growth, as a side effect, in a fixed amount of time.

The company wanted to study algae growth in natural water reservoirs, so they identified 15 different lakes suitable for the study. One lab was built nearby each lake beforehand. The company had requested that each lab study only ONE particular phosphoric salt and its effect on algae growth. An indoor pool was set-up in each lab to control for other potential variables, and water from the lake was drawn to fill the pools. The sizes of the pools and volumes of water from the lakes were identical across all pools and labs. Each pool in the study was built to have a surface area of 100 square meters.

You have visited each lab and checked whether and how much of the pool surface area was covered with algae before the study began. A particular type of phosporic salt was then

administered via aerial spray over the pool. Naturally, algae grow at different rates, depending on various conditions. This means that most likely the amount of algae growth will vary across different labs. Critically, however, the conditions WITHIN each lab were kept constant, so that such variation would affect the pool to the same extent before and after the treatment period, and that any change to the amount of algal bloom in the pool would have to be due to the type of phosporic salt administered. After two weeks, the surface area of the pool that was covered with algae was re-measured.

Your task is to go through records from the 15 labs. Each record contains information on the surface area covered by algae in the pool before and after treatment with the phosphoric salt in question. For each record, we are asking you to consider how strongly the chemical causes algae growth as a side effect.

To do so, we are asking you to imagine another lab in a different location that also has a pool inside. We are then asking you to imagine how much of the area of the pool would be covered with algae, once the chemical substance would be administered. We will ask you this question twice, once for a pool that has the same size as the one in the study, and once for a different sized pool.

A.7 Cover Stories For Experiment 3.6

Imagine you are a scientist conducting research on the runniness of engine oil. Engine oil is used to lubricate the moving parts of various internal combustion engines and increases engine performance and efficiency by creating a thin film between the surfaces of moving parts, hence reducing friction and energy loss via heat. Because of this function, the runniness of engine oil is crucial: Oil that is too thin may not properly stick on the surface of engine parts, whereas oil that is too thick may hinder the movement of the parts.

A company was interested to study the effects of 15 additives on the runniness of a particular type of engine oil. The company investigated the runniness of oil through a Free Flow Test procedure. In this procedure, 5 grams of oil are deposited on one end of a 10 cm long test slate, slanted at an angle of 45 degrees. After an interval of 5 minutes the total length (out of 10 cm) travelled by the drop of oil is measured as an indicator of runniness: The greater the distance the oil travelled, the greater the runniness of oil. For each additive, the scientists always tested the runniness of oil before adding the additives, and then mixed an additive into the oil and repeated the test.

The company has asked you to evaluate how strongly different additives in influencing the runniness of oil. To do this, you will see the results of the Free Flow Tests as described above. The results contain information on the length (out of 10 cm) of oil flow before and after adding each additive, for one additive at a time. Because different additives were tested in different laboratories and on different days, variations in climate (air temperature and humidity) might have affected the length of oil flow in the absence of any additive treatment between different tests.

To evaluate the effect of a particular additive on the runniness of engine oil, we will then ask you to imagine a new Free Flow Test with the same engine oil and the additive under investigation. We will ask you to predict the length of oil flow if the additive were administered to the oil. [For the higher limit condition, additionally, I included this sentence: We will ask this twice, once using a slate with the same length as used in the above investigation, and once with a longer slate.]

A.8 Cover Stories For Experiment 3.7

Imagine that you are a pharmaceutical consultant conducting research on the effectiveness of ointments in treating skin rash. Skin rash is a change to the skin causing it to become reddish, bumpy, itchy and sometimes painful. There are various identified causes of skin rash with food allergy being one of the most common. During an allergic reaction, the immune system reacts to certain proteins in food as if they were harmful substances. The body fights back by producing histamine, a chemical that is responsible for the skin rash.

Once triggered, skin rash can spread over large areas of skin. Administering ointment can reduce the affected area by alleviating the allergic reaction. Some ointments are more effective in reducing the spread than others. The purpose of this study is to investigate the effectiveness of various ointments in preventing rash spread on the skin.

In this study, there were 15 different labs participating, and each studied ONE particular ointment. To make the study consistent across the labs, only patients with similar skin rash reaction were selected for the study. Previous records of every patient have shown that after exposing themselves to certain seafood and in absence of any preventive measure, rash on their back would take up 10 centimeters square within an hour.

In each lab, one patient was exposed to the allergic food and the area of rash was measured one hour later. The particular ointment was then applied to the rash area and a second area measurement was made after an hour of exposure.

Your will see the records from the participating 15 labs, each investigating one of the 15 types of ointments. For each ointment, we are asking you to consider how effective you think it is in reducing skin rash. To do so, we are asking you to imagine a new patient with a food allergy

who is suffering from a rash. We are then asking you to imagine how much the skin rash area would be reduced, once the ointment would be administered.

A.9 Cover Stories For Experiment 3.9

Algae are a very common group of microplants that can be found in most aquatic systems. Algae, like other plants are autotrophs: they are capable of converting sunlight into food to stay alive. Algae can reproduce in either sexual or asexual, or both ways, depending on the environmental conditions they are in. Environmental factors, such as temperature, inorganic chemical nutrients, and salinity, regulate not only the method, but also the rate of reproduction.

High algae reproduction rates can result in a phenomenon known as algal bloom, in which algae concentrations may reach millions of cells per milliliter. In this state, the production of phytoplankton, as the base form of the marine food chain, is disrupted, hence endangering the whole ecosystem. Algal bloom is of concerns to many environmentalists. Consequently, there is now a research drive on how to manipulate the chemical composition of water bodies to prevent algal bloom. The best way to study algal bloom initially is an indoor setting. This allows researchers to keep atmospheric and physical conditions constant while investigating different water treatments.

Imagine that you are a biochemist investigating the influence of chemical substances on algae reproduction. Your investigation strives to prepare a chemical environment to prevent algae from growing and maturing in a fixed amount of time.

In this project, 15 different labs have participated, and each lab studied ONE particular chemical substance. Two indoor pools were set-up in each lab to control for other potential variables. The sizes of the pools and volumes of water filled in were identical in all pools and labs. Each pool in the study was built to have a surface area of 100 square meters.

In each lab, both pools were seeded with algae cysts. The first pool was retained as a control, whereas the chemical substance was then administered via aerial spray over the second pool. The two pools were in separate sections of each lab to ensure that the treatment could not affect the control pool. Other than the treatment though, the conditions in both pools were identical (i.e. chemical, biological, and physical characteristics of water and surrounding air, temperature, etc.). Naturally, algae mature and reproduce at different rates, depending on various conditions. This means that most likely the amount of algae growth will vary across different labs. Critically, however, the conditions WITHIN each lab were kept constant, so that such variation would affect both the control and treatment pools to the same extent. After two weeks, the area of water surface covered with matured algae in each pool was measured.

Your task is to go through records from the 15 participating labs. Each record contains information on the surface area covered by matured algae in the pools treated with the chemical in question, and the control pool without treatment. For each record, we are asking you to consider how strongly the chemical prevents algae reproduction.

To do so, we are asking you to imagine a new pool seeded with algae cysts. We are then asking you to imagine how much of the area of the pool would be covered with matured algae, once the chemical substance would be administered.

A.10 Cover Stories For Experiment 3.10

Imagine you are a scientist conducting research on the runniness of engine oil. Engine oil is used to lubricate the moving parts of various internal combustion engines and increases engine performance and efficiency by creating a thin film between the surfaces of moving parts, hence reducing friction and energy loss via heat. Because of this function, the runniness of engine oil is

crucial: Oil that is too thin may not properly stick on the surface of engine parts, whereas oil that is too thick may hinder the movement of the parts.

A company was interested to study the effects of 15 additives on the runniness of a particular type of engine oil. The company investigated the runniness of oil through a Free Flow Test procedure. In this procedure, 5 grams of oil are deposited on one end of a 10 cm long test slate, slanted at an angle of 45 degrees. After an interval of 5 minutes the total length (out of 10 cm) travelled by the drop of oil is measured as an indicator of runniness: The greater the distance the oil travelled, the greater the runniness of oil. For each additive, the scientists always tested the runniness of oil before adding the additives, and then mixed an additive into the oil and repeated the test.

The company has asked you to evaluate how strongly different additives in influencing the runniness of oil. To do this, you will see the results of the Free Flow Tests as described above. The results contain information on the length (out of 10 cm) of oil flow before and after adding each additive, for one additive at a time. Because different additives were tested in different laboratories and on different days, variations in climate (air temperature and humidity) might have affected the length of oil flow in the absence of any additive treatment between different tests.

To evaluate the effect of a particular additive on the runniness of engine oil, we will then ask you to imagine a new Free Flow Test with the same engine oil and the additive under investigation. We will ask you to predict the length of oil flow if the additive were administered to the oil. [For the higher limit condition, I included the following additional sentence: We will ask this twice, once using a slate with the same length as used in the above investigation, and once with a longer slate.]

A.11 Cover Stories For Experiment 4.1 and 4.5

Imagine you are a biochemist who has been hired by a multinational company to conduct research into algae reproduction. Algae have been getting more attention as an alternative source for biofuel production, but at the same time rapid algae growth could also throw local ecosystems out of balance.

The company was interested in investigating natural minerals to manage algae growth in 4 different climatic zones of the world - tropical, arid, mediterranean, and alpine. Within each zone, 2 locally occurring natural minerals were identified, and 2 different lakes were selected as study locations.

At each location, the company built a pool inside a lab filled with water drawn from the lake. All pools were built to have a surface area of 100 square meters. Research teams in each climate zone conducted experiments to test the influence of the 2 minerals local to that zone.

You will see results of the experiments, for all locations in each zone, containing information on the surface area (out of 100 square meters) covered by algae before and after treatment with the minerals. Within each climate zone, the two study locations were several hundred miles apart, and may have had different microclimates. This means that the surface area covered by algae at the beginning of the experiment (i.e. before the mineral treatment was administered) may vary between locations.

Your task is to go through the experimental records and make judgment on how ONE of the minerals influences algae growth. That mineral could cause growth, inhibit growth, or have no influence on algae growth.

A.12 Cover Stories For Experiment 4.2 and 4.6

Imagine that you are a materials scientist conducting research into the runniness of engine oil. Engine oil is used to lubricate the moving parts of various internal combustion engines and increases engine performance and efficiency by creating a thin film between the surfaces of moving parts, hence reducing friction and energy loss via heat. Because of this function, the runniness of engine oil is crucial: Oil that is too thin may not properly stick on the surface of engine parts, whereas oil that is too thick may hinder the movement of the parts.

Various additives can be used to manipulate the runniness of engine oil. Your company wanted to study the effects of various additives on four different varieties of engine oil - Bobil, Vatrol, Keszoil, and Ghotul.

For each oil, your company identified two chemical additives that they wanted to study. One department of the company investigated the influence of one chemical on each oil, while another department investigated the influence of both additives on each oil. Your task as the consulting scientist is to evaluate the combined results from both departments and indicate how ONE of the additives influences the runniness of each oil.

Each department tested the runniness of oil through a Drip-Test procedure, by measuring the splash area of 5 grams of oil dripped from 5 centimetres onto a test slate of 10 square centimetres. The greater the surface area on the test slate (out of the 10 square centimetres total) covered by oil, the greater the runniness of the oil. Department A tested each oil before adding any additives, and then mixed ONE additive into the oil and repeated the test. Department B also tested each oil before adding any additives, and then mixed TWO additives into the oil before repeating the test.

You will see results of the experiments from both departments, containing information on the splash area (out of 10 square centimetres) covered by oil before and after adding the chemicals, for one type of oil at a time. Because Departments A and B were in different geographical locations, differences in air temperature and humidity may have led to differences in the splash area covered by oil at the beginning of the experiment (i.e. before any chemicals were added).

Your task is to go through the experimental records and judge how ONE of the chemical additives influences the runniness of oil. That chemical additive could cause, prevent, or have no influence on the runniness of oil.

You can also select 'Cannot Tell' if you feel that the information provided does not allow an assessment of the influence of the chemical in question. Please use this option ONLY if you have a reason for WHY you cannot make a judgment. Do not use it just because you find the problem difficult. If you opt for 'Cannot Tell', you will have to explain in your own words why the information provided precludes an assessment of causal influence.

A.13 Cover Stories For Experiment 4.3 and 4.7

Imagine you are a biochemist who has been hired by a multinational company to conduct research into algae reproduction. Algae have been getting more attention as an alternative source for biofuel production, but at the same time rapid algae growth could also throw local ecosystems out of balance.

The company was interested in investigating natural minerals to manage algae growth in four different climatic zones of the world - tropical, arid, mediterranean, and alpine. Within each zone, two locally occurring natural minerals were identified and tested.

In each climate zone, the company identified two lakes where they studied algae growth. On the shore of each lake, the company built a research laboratory. Each laboratory housed two pools: one pool served as a control, and received no treatment, whereas the other pool received treatment.

Also, in each climate zone, the company decided that one laboratory was assigned to study the effect of ONE mineral, while the other was assigned to study the effects of BOTH minerals. Therefore, depending on which laboratory it was in, the treatment pool was treated with either one (Laboratory 1) or both (Laboratory 2) minerals.

All pools involved in this study were built to have a surface area of 100 square meters and filled with water drawn from their respective lakes.

You will see results of the experiments, for all locations in each zone, containing information on the surface area (out of 100 square meters) covered by algae for the pool treated with one or both minerals and also for the corresponding control pools without any treatment. Within each climate zone, the two lakes (and thus laboratories) were several hundred miles apart, and may have had different microclimates. This means that the surface area covered by algae in the absence of treatment may vary between locations.

Your task is to go study the experimental records and judge how ONE of the minerals influences algae growth. That mineral could cause growth, inhibit growth, or have no influence on algae growth.

A.14 Cover Stories For Experiment 4.4 and 4.8

Imagine that you are a materials scientist conducting research into the runniness of engine oil. Engine oil is used to lubricate the moving parts of various internal combustion engines and increases engine performance and efficiency by creating a thin film between the

surfaces of moving parts, hence reducing friction and energy loss via heat. Because of this function, the runniness of engine oil is crucial: Oil that is too thin may not properly stick on the surface of engine parts, whereas oil that is too thick may hinder the movement of the parts.

Various additives can be used to manipulate the runniness of engine oil. Your company wanted to study the effects of various additives on four different varieties of engine oil - Bobil, Vatrol, Keszoil, and Ghotul.

For each oil, your company identified two chemical additives that they wanted to study. One department of the company investigated the influence of one chemical on each oil, while another department investigated the influence of both additives on each oil. Your task as the consulting scientist is to evaluate the combined results from both departments and indicate how ONE of the additives influences the runniness of each oil.

Each department tested the runniness of oil through a Drip-Test procedure, by measuring the splash area of 5 grams of oil dripped from 5 centimetres onto a test slate of 10 square centimetres. The greater the surface area on the test slate (out of the 10 square centimetres total) covered by oil, the greater the runniness of the oil. Both departments tested two samples of each oil. One sample consisted of the original oil without any additives, and the other sample was mixed with one additive in Department A, and both additives in Department B.

You will see the results of the experiments from both departments, containing information on the splash area (out of 10 square centimetres) covered by oil from the control sample, and from the sample that was mixed with chemical additives, for one type of oil at a time. Because Departments A and B were in different geographical locations, differences in air temperature and humidity may have led to differences in the splash area covered by oil in the control group (i.e. oil that was not added any chemical).

Your task is to go through the experimental records and judge how ONE of the chemical additives influences the runniness of oil. That chemical additive could cause, prevent, or have no influence on the runniness of oil.

You can also select 'Cannot Tell' if you feel that the information provided does not allow an assessment of the influence of the chemical in question. Please use this option ONLY if you have a reason for WHY you cannot make a judgment. Do not use it just because you find the problem difficult. If you opt for 'Cannot Tell', you will have to explain in your own words why the information provided precludes an assessment of causal influence.

A.15 Cover Stories For Experiment 4.9

Imagine you are a biochemist who has been hired by a multinational company to conduct research into algae reproduction. Algae have been getting more attention as an alternative source for biofuel production, but at the same time rapid algae growth could also throw local ecosystems out of balance.

The company was interested in investigating natural minerals to manage algae growth in seven different climatic zones of the world - tropical, arid, mediterranean, alpine, tundra, savanna, and steppe. Within each zone, 2 locally occurring natural minerals were identified, and 2 different lakes (or natural water reservoirs) were selected as study locations.

At each location, the company built a pool inside a lab filled with water drawn from the lake. All pools were built to have a surface area of 100 square meters. Research teams in each climate zone conducted experiments to test the influence of the 2 minerals local to that zone.

You will see results of the experiments, for all locations in each zone, containing information on the surface area (out of 100 square meters) covered by algae before and after treatment with the minerals. Within each climate zone, the two study locations were several

hundred miles apart, and may have had different microclimates. This means that the surface area covered by algae at the beginning of the experiment (i.e. before the mineral treatment was administered) may vary between locations.

Your task is to go through the experimental records and make a judgment on how ONE of the minerals influences algae growth. That mineral could cause growth, inhibit growth, or have no influence on algae growth.

You can also select 'Cannot Tell' if you feel that the information provided does not allow an assessment of the influence of the mineral in question. Please use this option ONLY if you have a reason for WHY you cannot make a judgment. Do not use it just because you find the problem difficult. If you opt for 'Cannot Tell', you will have to explain in your own words why the information provided precludes an assessment of causal influence.

A.16 Cover Stories For Experiment 4.10

Imagine you are a biochemist who has been hired by a multinational company to conduct research into algae reproduction. Algae have been getting more attention as an alternative source for biofuel production, but at the same time rapid algae growth could also throw local ecosystems out of balance.

The company was interested in investigating natural minerals to manage algae growth in seven different climatic zones of the world - tropical, arid, mediterranean, alpine, tundra, savanna, and steppe. Within each zone, two locally occurring natural minerals were identified and tested.

In each climate zone, the company identified two lakes where they studied algae growth. On the shore of each lake, the company built a research laboratory. Each laboratory housed two

pools: one pool served as a control, and received no treatment, whereas the other pool received treatment.

Also, in each climate zone, the company decided that one laboratory was assigned to study the effect of ONE mineral, while the other was assigned to study the effects of BOTH minerals. Therefore, depending on which laboratory it was in, the treatment pool was treated with either one (Laboratory 1) or both (Laboratory 2) minerals.

All pools involved in this study were built to have a surface area of 100 square meters and filled with water drawn from their respective lakes.

You will see results of the experiments, for all laboratories in each zone, containing information on the surface area (out of 100 square meters) covered by algae for the pool treated with one or both minerals and also for the corresponding control pools without any treatment. Within each climate zone, the two lakes (and thus laboratories) were several hundred miles apart, and may have had different microclimates. This means that the surface area covered by algae in the absence of treatment may vary between laboratories.

Your task is to go through the experimental records and make a judgment on how ONE of the minerals influences algae growth. That mineral could cause growth, inhibit growth, or have no influence on algae growth.

You can also select 'Cannot Tell' if you feel that the information provided does not allow an assessment of the influence of the mineral in question. Please use this option ONLY if you have a reason for WHY you cannot make a judgment. Do not use it just because you find the problem difficult. If you opt for 'Cannot Tell', you will have to explain in your own words why the information provided precludes an assessment of causal influence.

Appendix B: Determination of Bin Size for Histogram Analysis in Chapter 3

B.1 Determination of Bin Size for Histogram Analysis in Chapter 3

When developing histogram, the size of bin is vital as it influences the histogram itself, as well as any derived conclusion. A bin size that is too big would de-sensitive certain judgment pattern and possibly amalgamated two or more judgment patterns into one; likewise, a bin size that is too small would split supposedly a judgment pattern into two or more pattern. Consequently, if the bin size was either too big or too small, it would render different conclusion, as to what it supposed.

Thus, for this analysis, I defined the bin size by determining the minimal gap between the predictions of proportion and difference strategies. To do this, I considered all predictions of the 15 conditions used in the experiments, for both strategies (see Table B.1). I computed these predictions using the base rate of 25, and upper limit of 100, for generative scenario. This were the settings of the cover story in the experiments.

From these predictions, I calculated prediction gaps for these settings by subtracting every prediction of the proportion strategy (14 items) with every prediction of the difference strategy (5 items). The result was a 14 x 5 matrix consisted of prediction gaps between the proportion and difference strategy (see Table B.2). To determine bin size for the histogram analysis, I needed to identify the gap between these strategies' predictions so that the histogram could properly differentiate the judgments to either reflect the proportion of difference strategy: In other words, I needed to identify the smallest gap. From the matrix, the smallest gap was 0.00. This value, however, is not suitable to be used as the bin size because it would render the histogram into vertical lines on all values. In addition, the gap 0.00 indicates that the predictions

between the two strategies are the same/overlapping. Therefore, the smallest non-zero value of 6.25 would be more appropriate as the bin size. To further simplify, and to be more cautious, I declared the bin size to be 5 (i.e. 1.25 smaller than the smallest non-zero gap).

The bin size of 5, as in the above explanation, was for analysing data of experiment with upper limit of 100. In Chapter 3, the histogram analysis was used in experiments involving upper limit of 500 (i.e. higher limit condition), as well as 10 and 50 in other experiments. Thus, for these upper limits, I repeated the same procedure to determine bin size for each limit (see Table B.3 for summary of bin sizes for all of these limits).

Table B.1: Predictions of 15 experiments conditions for both the proportion and difference strategies

Conditions		Predictions (base rate = 25, limit = 100)	
Q(e|c)	Q(e|¬c)	Difference	Proportion
1.00	1.00	0.00	-
0.75	0.75	0.00	0.00
0.50	0.50	0.00	0.00
0.25	0.25	0.00	0.00
0.00	0.00	0.00	0.00
1.00	0.75	25.00	75.00
0.75	0.50	25.00	37.50
0.50	0.25	25.00	25.00
0.25	0.00	25.00	18.75
1.00	0.50	50.00	75.00
0.75	0.25	50.00	50.00
0.50	0.00	50.00	37.50
1.00	0.25	75.00	75.00
0.75	0.00	75.00	56.25
1.00	0.00	100.00	75.00

Table B.2: Prediction Gaps Matrix

Predictions		Difference				
		0	25	50	75	100
Proportion	0.00	0.00	25.00	50.00	75.00	100.00
	0.00	0.00	25.00	50.00	75.00	100.00
	0.00	0.00	25.00	50.00	75.00	100.00
	0.00	0.00	25.00	50.00	75.00	100.00
	75.00	75.00	50.00	25.00	0.00	25.00
	37.50	37.50	12.50	12.50	37.50	62.50
	25.00	25.00	0.00	25.00	50.00	75.00
	18.75	18.75	6.25	31.25	56.25	81.25
	75.00	75.00	50.00	25.00	0.00	25.00
	50.00	50.00	25.00	0.00	25.00	50.00
	37.50	37.50	12.50	12.50	37.50	62.50
	75.00	75.00	50.00	25.00	0.00	25.00
	56.25	56.25	31.25	6.25	18.75	43.75
	75.00	75.00	50.00	25.00	0.00	25.00

Table B.3: Bin sizes for all limits

Limit	Bin size
10	0.5
50	2.5
100	5
500	25

B.2 Determination of Gamut for Tendency Analysis in Chapter 3

Tendency analysis determined which strategy prediction participants' scores fell into. Because participants' scores did not exactly match the prediction values, I need to define a gamut for every prediction value, of which if the scores were within this range, they would be considered to match with the prediction.

To do this, I identified a condition that has predictions that do not overlap but very close to one another: Figure 3.2 shows that the generative predictions for condition [0.75:0.25] fit these criteria. In this condition, in particular, predictions for the difference, proportion strategy, and base rate neglect are very close. Then, I determined the gap between these predictions as

follow: difference (0.50) vs. proportion (0.667) = 0.167, and proportion (0.667) vs. base rate neglect (0.75) = 0.083. The smallest gap is between the proportion strategy and base rate neglect.

Using this information, I defined the gamut for tendency analysis as ± 4.17% (i.e. half of 0.083 gap) from a prediction value. This is to avoid range overlapping of a prediction to another. For instance, if a prediction value is 0.50, then its gamut begins at 0.46 until 0.54. This rule, however, is different for ceiling and floor prediction values. For prediction values of ceiling (1) and floor (0), the gamut does not stretch out of this. In other words, a prediction of 1 would have a range beginning at 0.96 until 1, and a prediction 0 would have a range from 0 until 0.04.

Appendix C: Multiplication versus Proportion Strategy

The main idea of the proportion strategy entails reasoning about the independent influence of candidate cause and background causes on the effect magnitude. Computationally, reasoners adopting the proportion strategy would need to initially determine the opportunity for effect to change up to its maximum magnitude possible; this involves considering the upper limit and the base rate of the effect magnitude. Only then, he or she could calculate how much from the opportunity the effect magnitude had actually changed.

Unlike the proportion strategy, which is sensitive to the upper limit, the multiplication strategy stems from an interaction between the candidate and background causes. Computationally, multiplication-based reasoners determine effectivity of candidate cause onto the effect by initially consider the effect magnitude the background causes produce, i.e. the base rate. Then, he or she calculated how many times more the candidate cause produced the effect with respect to the base rate.

Computationally, both the proportion and multiplication strategies involve finding out causal index by computing ratios. See Figure C.1 for visualisation of both strategies. In this figure, proportion index, $prop = \frac{x}{y}$ in both generative and preventive scenario. In contrast, multiplication index, $mult = \frac{x}{z}$ in generative scenario, but $mult = \frac{z}{y}$ in preventive scenario.

The difference between the proportion and multiplication strategy lies in the denominator. The inconsistency between these denominators is the key in explaining the asymmetry of both strategies' predictions in preventive and generative scenarios. Recall that predictions of these two strategies in experiments of Chapter 2 and 3 were overlapping only in preventive, but distinct in generative scenario. This is because, those denominators entails information about reasoning references of both strategies.

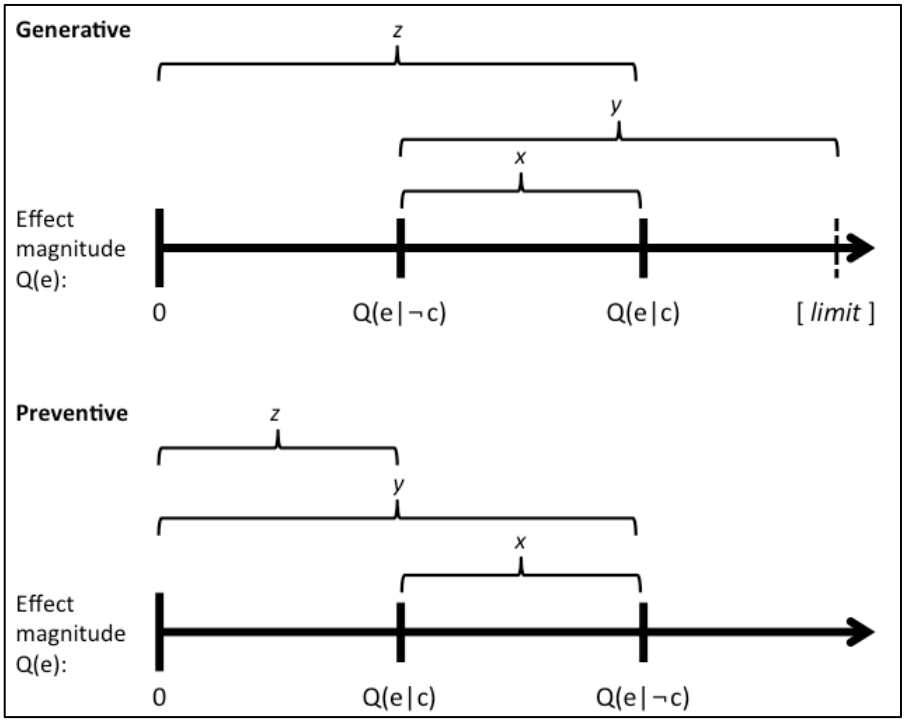


Figure C.1. Visualisation of computing the proportion and multiplication strategies. The limit on the top part of the figure corresponds to artificial limits used in the experiments.

In preventive scenario, both strategies anchor their denominator on the same reference, i.e. *y* in Figure C.1. On the contrary, in generative scenario, while the proportion strategy retains the same anchors on *y*, the multiplication strategy switched to *z*. In other words, in preventive scenario, both strategies adopt the same reference, which is the limit of 0; whereas in generative scenario, while the proportion strategy uses the upper limit as the reference, the multiplication strategy considers the origin i.e. 0 as its reference.

Table C.1: Comparison of proportion and multiplication indices in both scenarios

	Proportion	Multiplication
Generative	$\frac{x}{y}$	$\frac{x}{z}$
Preventive	$\frac{x}{y}$	$\frac{z}{y}$

Table C.1 summarises this. In short, because of this, the predictions of the proportion and multiplication only overlap in preventive scenario, and not in generative.