2020

Theory protection in human associative learning and formally representing uncertainty about novel stimuli

Spicer, Stuart Gordon

UNIVERSITY OF PLYMOUTH

THEORY PROTECTION IN HUMAN ASSOCIATIVE LEARNING

AND

FORMALLY REPRESENTING UNCERTAINTY ABOUT NOVEL STIMULI

Stuart G. Spicer

A thesis submitted to the University of Plymouth in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Psychology

January 2020

Acknowledgements

Firstly, I would like to thank the School of Psychology at the University of Plymouth for providing me with the funding and resources that enabled me to complete this project.

I would especially like to thank both of my supervisors, Dr Peter Jones and Professor Andy Wills, for their support and guidance throughout my PhD. They have both offered a wealth of experience and expertise, allowing me to learn and develop my skills and knowledge throughout the project.

A big thank you goes to Professor Chris Mitchell, for his support and encouragement throughout my PhD. He has provided a wonderful source of inspiration and insight as I have developed my theoretical ideas.

I would like to give further thanks to Dr Mark Haselgrove, who’s input into my earlier career helped me to develop an interest in associative learning, and some of the issues discussed within my thesis.

Additional thanks goes to two project students. Firstly, Lenard Dome for his input into the formal model implementations used for the simulations in Chapter 4, and secondly Katie Blake for her help with data collection in the final experiment in Chapter 3.

Finally, I would like to thank my family, friends, and colleagues for their ongoing support throughout my PhD.

Author’s declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment. Relevant scientific seminars and conferences were regularly attended at which work was presented. Several papers were prepared and submitted for publication:

Spicer, S. G., Mitchell, C. J., Wills, A. J., and Jones, P. M. (2019). Theory protection in associative learning: humans maintain certain beliefs in a manner that violates prediction error. *Journal of Experimental Psychology: Animal Learning and Cognition*.

Spicer, S. G., Mitchell, C. J., Wills, A. J., Blake. K. L., and Jones, P. M. (under review). Theory protection: do humans protect existing associative links? *Journal of Experimental Psychology: Animal Learning and Cognition*.

Spicer, S. G., Mitchell, C. J., Wills, A. J., Dome, L., and Jones, P. M. (under review). Representing uncertainty in the Rescorla-Wagner model: blocking, the redundancy effect, and outcome base rate. *Open Journal of Experimental Psychology and Neuroscience.*

Word count for main body of thesis: 38,396

Signed:

Date:

Abstract

Three collections of data are presented in this thesis, with the broad aim of investigating a potential role for theory protection in human associative learning. According to theory protection, people should resist updating their existing knowledge (i.e. resist new learning), even when faced with evidence that contradicts what they already know. In other words, people should maintain established associations between environmental cues and outcomes wherever possible. Theory protection differs from typical prediction error accounts of learning (e.g. Bush & Mosteller, 1951; Rescorla & Wagner, 1972; Rescorla, 2001). According to prediction error accounts, people should update existing associations (i.e. learn) most readily when the outcomes they encounter are most discrepant with what they predict. Details about these accounts are introduced in Chapter 1, along with several other theories and phenomena that are central to the subsequent chapters. In the first set of experiments (reported in Chapter 2), human participants were initially trained with a set of cues, each of which was followed by the presence or absence of an outcome. In a subsequent training stage, two of these cues were trained together, and the amount learned about each of the cues was compared, using a procedure based on Rescorla (2001). In each experiment, the cues differed in both their prediction error (with respect to the outcome), and the confidence participants should have about their causal status. The cue with the larger prediction error was always the cue with lower confidence about its causal status. In an apparent violation of prediction error, participants always learned more about the cue with the smaller prediction error, supporting a theory protection account of learning. Participants appeared to protect their existing beliefs about cues with a known causal status, instead attributing unexpected outcomes to cues with a comparatively ambiguous causal status. The second set of experiments (reported in Chapter 3) provides further evidence of

theory protection, except that the cues, outcomes and experimental scenario differed to those in the Chapter 2 experiments. Chapter 3 also includes direct testing of the theory protection account against the predictions of both Pearce and Hall's (1980), and Mackintosh's (1975) attentional accounts of learning. The results were not consistent with either attentional theory. The final set of data (reported in Chapter 4) includes the results of formal model fitting simulations. The findings illustrate a simple way of representing participants' lack of confidence about the causal status of novel cues. This was achieved by allowing the initial strength of associations (between cues and outcomes) to be an intermediate value. Importantly, the best fitting initial associative strength was shown to change in line with the overall proportion of trials on which the outcome occurs. Free and open resources to support formal modelling in associative learning are briefly introduced. Chapter 5 provides a general discussion of all the findings, whilst also setting out a future programme of research, so that the theory protection account can be developed into a formal model of human associative learning.

Contents

List of figures

1: General Introduction

The primary aim of the research reported in this thesis was the investigation of theory protection in human associative learning. Associative learning refers to the formation of of mental links between phenomena in the environment. This includes rats learning that specific behaviours will result in a reward (e.g. Skinner, 1938), or humans learning that certain people are associated with either positive or negative attributes. In the context of this thesis, these associations are specific to learning about causation (i.e. a specific event causing a specific outcome). Theory protection encapsulates the notion that people resist updating established beliefs, even when faced with contradictory evidence. Once such beliefs have been learned, people should protect them as much as possible, changing their views in the most minimally available way (e.g. Harman, 1986). Beliefs are conceptually rather a broad area, but in the context of this thesis, they specifically refer to knowledge about the causal status of environmental phenomena, with respect to associated outcomes. For example, you might hold a belief that eating eggs causes stomach ache.

In several of the experiments reported in this thesis, people appeared to protect their beliefs, when faced with seemingly contradictory evidence. The experimental participants instead attributed the occurrence of this unexpected evidence to other readily available causes. Specifically, participants protected their beliefs about stimuli with a known causal status, instead attributing an unexpected event to stimuli about which no strong beliefs were held. Of course, protecting beliefs and attributing the occurrence of unexpected evidence elsewhere is by no means a phenomenon unique to these experiments. This type of process has been demonstrated in research into

prejudice and attribution error (e.g. Hewstone, 1990; Hewstone, Rubin, & Willis, 2002), where positive behaviour by members of an outgroup is attributed to chance factors, rather than being attributed to the inherent characteristics of that outgroup itself. These attributions allow a negative opinion of outgroups to be maintained, contributing to prejudice and bias.

In the context of associative learning, theory protection should prevent new learning from occurring, once associations have been formed. More informally, theory protection represents the idea that people are somewhat stubborn and rigid in their beliefs, once they have learned about something. Theory protection contrasts with the established notion that learning is governed by prediction error (e.g. Rescorla, 2001). Prediction error is a formal way of representing the discrepancy between what is expected and what occurs, when an event is unexpected or surprising. According to prediction error accounts of learning, people should learn the most when what they expect contrasts most with what they experience. Consequently, a larger prediction error should result in a larger updating of existing beliefs. To illustrate how these processes operate, and how they differ, imagine that you have been enjoying a long running box-set TV show on your favourite streaming service. Over the course of the show, the quality of the episodes has been extraordinarily high, and you have therefore learned to associate this show with excellent quality writing, acting and directing. Let us assume that the first episode of the latest season airs. As a consequence of the high quality of the preceding seasons, you should predict that this episode will also be of a very high quality. However, contrary to what you predict, the first episode in fact turns out to be appalling.

According to a prediction error account, there will be a large discrepancy between what you predict (i.e. excellent quality) and what actually occurs (i.e. terrible quality). This large prediction error should drive a large amount of learning. Consequently, your expectations about the rest of the season should be considerably poorer. However, according to theory protection, you should resist updating your expectations about the rest of the season, particularly if there are other factors to which you can attribute the poor quality of the first episode. For example, the first episode could simply be setting the pieces in place for a fantastic final run of episodes, rather than being concerned with telling a quality story in itself. Alternatively, the disappointing first episode could simply be an uncharacteristic wobble in quality that is unlikely to be repeated. After viewing this first episode, your expectations for the rest of the season should remain relatively unchanged and you should anticipate that higher quality episodes will follow.

However, as more episodes of the season air, they also turn out to be terrible. According to a prediction error account, there should be less learning as the season progresses, as you will already predict a poorer quality set of episodes. There will no longer be a large prediction error to drive learning, so there will be less scope for any new updating of your beliefs. However, according to theory protection, you should continue to resist updating your expectations about the rest of the season, desperately holding on to the hope of some brilliant quality TV that the writers will pull out of the bag before the season finale. Eventually, of course, the weight of evidence will become too great and you will be forced to accept the overwhelming reality of the situation. However, you may continue to look for factors to which you can attribute the failure of this now-tarnished show, into which you have invested so many of your precious hours.

Theory protection and prediction error and both involve more complex processes than this illustrative example, but the important point is that these differing accounts can lead to different learning outcomes in certain situations. Consequently, these processes can be tested against each other, with the goal of understanding which one is the better account of human associative learning. Whilst this was the overriding goal of the current research, a further goal was to modify existing formal models of human associative learning, in a way that allows for people's uncertainty about the causal status of novel cues to be mathematically represented. Over the course of this brief introductory chapter, theory protection and prediction error are unpacked in more detail. Some of the complexities alluded to above are also discussed in the subsequent chapters. Additionally, this chapter incorporates a brief summary of what else will be covered in the subsequent chapters of this thesis.

Prediction error is a core component of many influential models of associative learning (e.g. Bush & Mosteller, 1951; Mackintosh, 1975; Rescorla & Wagner, 1972). As briefly outlined above, such theories state that learning occurs when humans and non-human animals encounter surprising outcomes (or events). There are many scenarios in which prediction error is proposed to govern learning. These scenarios will necessarily involve some kind of discrepancy (i.e. error) between a predicted outcome and an experienced outcome. To give another example, if you expect that a certain medication (i.e. cue) will result in the side effect of a headache (i.e. outcome), but no headache occurs, then there will be a large error between what you predicted and what occurred. Consequently, your expectations about that medication will change and your future predictions of headache after consumption of the medication will be reduced. Eventually, after further use of this medication, you will be able to accurately predict that it is safe to use. Alternatively, if a

person expects that a type of food is safe to eat, but they experience an allergic reaction after eating that food, then learning will take place and expectations about that food will update accordingly. Since learning is proportional to prediction error, learning ceases once the prediction error is eliminated. At this point, learning is said to have reached asymptote. Much of the research supporting prediction error as a determinant of learning stems from non-human animal conditioning research (e.g. Pavlov, 1927; Kamin, 1969; Rescorla, 2001). All other things being equal, a larger prediction error should always result in greater learning.

Evidence for learning being proportional to prediction error is well established and has a long history. One of the clearest demonstrations of this relationship comes from "learning curves", where the reduction in learning rate, as prediction error reduces, can be plotted as a decelerating curve. For example, rats might be experimentally trained that pressing a lever results in the reward of a food pellet. As they learn, their delay before pressing the lever will reduce over time. However these reductions in time will become incrementally smaller until there is no further reduction. This type of learning is known as acquisition, and the plottable curve is known as an acquisition curve. Similarly, if the lever press ceases to result in a reward, the rats will also learn about this. This type of learning is known as extinction. A plottable extinction curve will be produced, as the delay before pressing the lever increases; rapidly at first, before slowing and settling at a level representing chance. Demonstrations of acquisition and extinction curves can be seen in examples such as Skinner (1938), Graham and Gagne (1940), Guttman (1953), and Lewis (1956).

In a study with rats and pigeons, Rescorla (2001) demonstrated a greater change in responding (signifying greater learning) for cues with a larger prediction error. In one experiment rats were shown two cues (e.g. a light and a sound) simultaneously (i.e. in a compound). Both of these cues were already familiar to the rats. One of these cues (which will be referred to as cue A) previously resulted in the occurrence of an outcome (e.g. a food pellet), while the other cue (which will be referred to as cue B) resulted in the absence of this outcome (e.g. no food pellet). If + and – are used to represent the presence or absence of the outcome, respectively, then the rats experienced A+ B- training prior to experiencing the compound of these cues. When the rats subsequently experienced this compound, the outcome was absent (AB-). The results indicated that the rats learned more about A, supporting a prediction error account. However, this seems at odds with what one might expect in human causal learning. During A+ trials participants would presumably form a belief that A causes the outcome. On the other hand, B should be somewhat ambiguous after B- training because there is no information available to indicate whether B is neutral or preventative (i.e. inhibitory), with respect to the outcome. To produce a result that is analogous to Rescorla's, participants would need to change their existing beliefs about A rather than attributing a surprising outcome to B, which is ambiguous. Perhaps a more intuitive prediction would be for human participants to maintain their belief (i.e. protect their theory) about the causal status of A, and instead attribute the absence of the outcome (during the AB- trials) to B, resulting in more learning about B. If true, this would run contrary to Rescorla's proposal.

In order to explain how theory protection and a prediction error account can be tested against each other, there is a little more complexity to Rescorla's theoretical and

methodological approach that needs to be discussed. Importantly, not all kinds of prediction error are the same. According to Bush and Mosteller's (1951) theory, learning about each cue is determined by the discrepancy between the outcome that occurs and the outcome that was predicted by that cue alone. This is referred to as individual prediction error. However, following the observation that knowledge about one cue can interfere with learning about another (e.g. the blocking effect, Kamin, 1969), Rescorla and Wagner (1972) suggested that learning was instead governed by overall prediction error; that is, the discrepancy between the outcome that occurs and an aggregate prediction derived from all the cues that are present. It is not necessary to know the equations for these two prediction error models for the majority of this thesis, but they are included in the introduction to Chapter 4 (4.1). Subsequently, Rescorla (2001) provided evidence that learning is governed by a mixture of individual prediction error and overall prediction error. In other words, when more than one cue is present, learning can be greater (but not smaller) for those cues whose individual causal status is most discrepant from the outcome. It is also necessary to explain the rest of the design of the Rescorla (2001) rat experiment described above, so that the design of the experiments in Chapters 2-3 make sense. In this experiment, the rats were initially trained with two excitatory (i.e. followed by the outcome) cues A+ and C+, and two non-reinforced (i.e. not followed by the outcome) cues B- and D-. After the completion of this initial training, the rats were then trained with a non-reinforced compound (AB-) containing one previously encountered excitatory cue A, and one previously encountered non-reinforced cue B. Following this, the amount learned about A and B on the AB- trials was compared using a compound testing procedure in which compounds AD and BC were presented.

This compound testing procedure provides a way of comparing the amount of learning for two cues that are trained from different starting associative strengths (i.e. strength of the association between a cue and an outcome). This is required because the relationship between associative strength and responding is not necessarily linear (Gluck & Bower, 1988), which makes it difficult to assess the amount of learning for individual cues. For example, an increase in associative strength from 0 to 0.5 (where 1 is the maximum value) would not necessarily translate into a change in responding from 0 to 5 on an 11-point response scale (where 10 is the maximum value), in the case of a human experiment. This same issue also applies to the kind of responses that non-human animals, such as rats, make in learning experiments. The change in responding could be either proportionally larger or proportionally smaller than the change in associative strength. In other words, rather than a 1:1 linear relationship, associative strength could instead map onto responding via a logistic function. However, as one value increases, the other value should also increase. Therefore, it would never be the case that an increase in associative strength should result in a decrease in responding. Rescorla's experiments circumvented the issue of non-linear mapping by using the above test compounds that each contained a cue that had previously been paired with the outcome (A or C) alongside a cue that had previously been paired with the absence of the outcome (B or D). Hence, in the absence of the second (AB-) stage, both test compounds would have resulted in equal responding. Any difference in responding to these compounds during the test must therefore have been the consequence of differences in learning about A and B on the AB- trials.

In the above (Rescorla, 2001) design, individual prediction error would lead to more learning about A during the compound training stage, since the outcome it predicts

following the initial training (A+) is the most discrepant with the outcome that occurs during the compound stage (AB-). Overall prediction error would result in equal learning about cues A and B, because they effectively ‘share’ their prediction error. This means that cues presented together account for an equal share in any change in associative strength that occurs from learning. The results showed less responding to the compound AD than BC, suggesting a greater decrease in associative strength for cue A than cue B. These results were taken by Rescorla to be evidence of individual prediction error. Had learning been solely based on overall prediction error, there would have been no difference in responding to the compounds at test. Rescorla’s (2001) proposal of a mixture of individual and overall prediction error was subsequently challenged by Holmes, Chan and Westbrook (2019), who showed that overall prediction error alone can account for these findings if the function mapping associative strength to responding is appropriately modified. As already outlined, the logic of this is that a non-linear mapping function allows equal changes in associative strength to result in unequal changes in responding, as well as unequal summation of individual cues into compounds.

In the case of humans, if participants protect their theory about A being a cause of the outcome during the AB- trials, the opposite result to Rescorla (2001) should be observed. This is because more should be learned about B, leading to reduced responding (e.g. lower causal ratings being assigned at test) to the BC compound. Participants should learn that B is preventative of the outcome, meaning that it would prevent C from causing the outcome. Even a modified response function (Holmes et al., 2019) should not permit such a finding under a prediction error account, since a cue with a smaller prediction error should not be learned about more than a cue with a larger

prediction error, assuming both cues are equally salient. In the case of two cues being trained in compound, the idea that the cue undergoing a larger change in associative strength would result in the smaller change in responding seems implausible. However, the view that human participants might maintain existing associations, in spite of a large prediction error, does seem plausible. In fact, such a view is consistent with theoretical approaches seen in other cognitive fields. For example, the suggestion that humans should resist updating their beliefs has parallels with theories about schemata (e.g. Bartlett, 1932), in terms of humans accommodating new information into existing frameworks. Also, the processing of new cue-outcome associations, in a manner that fits with existing causal associations, is comparable to confirmation bias (Wason, 1960). Additionally, attribution of an unexpected outcome to one cue, without beliefs about other cues needing to be updated, could be considered efficient in a manner that is similar to the concept of the cognitive miser (Fiske & Taylor, 1984). This latter concept in itself has obvious parallels with the minimalist updating of views (Harman, 1986). mentioned at the start of this chapter. It is worth briefly noting that this minimalist approach has been challenged by Walsh and Johnson-Laird (2009), who found evidence that some beliefs do not always update in the most informationally minimalist way. However, this updating of beliefs was based on participants explaining contradictory evidence thorough the use of disabling conditions. According to this process, learned beliefs (e.g. the striking of a match will cause it to light) are still protected, if exceptions to this can be attributed to disabling conditions (e.g. the match being soaking wet). Furthermore, such disabling conditions are themselves based on learned theories and beliefs about how things operate in our everyday environments, meaning that existing frameworks of beliefs are still maintained. Despite the influence of these theoretical approaches in fields such as decision making, there is currently no formal theory of

associative learning that implements this kind of process. A possible reason is the apparent lack of an obvious variable to govern learning. However, confidence about the causal status of cues provides one promising candidate. Specifically, if participants are more confident about the causal status of a cue, they should be more resistant to updating their beliefs (i.e. greater theory protection), resulting in less learning.

Whist there is no formal theory of associative learning that implements confidence in the manner described above, there are some phenomena and theories that encompass similar ideas. For example, latent inhibition (Lubow, 1973) is a phenomena concerning cues that have previously been encountered, without causing an outcome. If these previously encountered cues are subsequently trained as causing a specific outcome, they are learned about more slowly than novel cues that are given equivalent training. One common explanation of latent inhibition is that differences in the attention paid to cues, on the basis of their familiarity, alters the rate of learning (Mackintosh, 1975). In the case of humans, this phenomena could also be explained in terms of theory protection. If a cue has been demonstrated not to cause an outcome, then this belief would need to be updated, once it is subsequently trained with that outcome. However, latent inhibition has also been demonstrated in non-human animals. It could be that differing mechanisms produce equivalent effects in different species, but it may also be that this is not a demonstration of theory protection in humans.

Latent inhibition differs from standard inhibition (also referred to as conditioned inhibition). As briefly mentioned above, inhibition is learning that a specific cue (i.e. an inhibitor) is preventative of an outcome. For example a certain medication may actively prevent an allergic reaction from being caused by a type of food. Inhibition can be

tested for, using retardation of acquisition, in which a trained inhibitor and an excitatory cue are placed in a test compound. Reduced responding to this test compound provides evidence of the inhibitor preventing the outcome from being caused by the excitatory cue (e.g. Rescorla, 1971). Inhibition is relevant to the Rescorla (2001) experimental design, if applied to humans (and also to the experiments presented in Chapter 3). The A+ AB- design described earlier is known as a (type of) feature negative discrimination. There is evidence from rat experiments that cue B can acquire inhibitory control over subsequent responding, when tested with different excitatory cues (e.g. Holland & Coldwell, 1993). Initially, this might seem contradictory, given than Rescorla demonstrated that rats learn more about A. However, simple discriminations (without a compound design) do not allow for a comparison of the amount learned about cues A and B during the AB- trials. Even if A remains excitatory and B becomes inhibitory, this does not mean that B is learned about more than A. Rescorla's (2001) results support this notion. Inhibition is considered in greater detail in Chapter 3.

It is also worth briefly discussing theory protection in terms of context. Returning to the example of extinction, participants should initially resist altering what they have already learned about a cue. For example, if a specific food has been learned as causing stomach ache, and is subsequently trained as not causing stomach ache, participants should protect their belief that this food causes stomach ache. In the A+ AB- example, there is another readily available cue, to which the unexpected absence of the outcome can be attributed, allowing beliefs about the excitatory cue to be protected. However, another cue might not always be available. There is evidence that context plays an important role in such circumstances. Extinction appears to be context specific in non-human animals (e.g. Bouton, 1994; Bouton & Todd, 2014). For example, if a cue is learned as

an excitor in context 1 and then extinguished in context 2, the extinguished responding will return if that animal is placed back in context 1. Similarly, if a cue is both learned about and extinguished in context 1, the extinguished responding will return if the animal is placed in context 2. This type of context-dependent “renewal” has also been demonstrated with cues learned as inhibitors (Bouton & Todd, 2014). This process is comparable to the concept of theory protection in humans, in that an association is maintained (i.e. renews after extinction) in spite of new information being learned. However, the research cited above was conducted with non-human animals (e.g. rats). Given the Rescorla (2001) results, it is unlikely that theory protection is governing this process in rats. Bouton and Todd (2014) outline a plausible explanation, in which context can either have a direct effect on responding (i.e. through an inhibitory association between context and responding), or become associated with a cue-outcome association (i.e. an association with an association).

In the case of humans, context could still provide a way of explaining an unexpected outcome and protecting a theory. For example, if you expect that a certain type of food is safe to eat, but you experience an allergic reaction while eating at a new restaurant, you might protect your belief that the food is safe to eat, and attribute the allergic reaction to the context of being in that restaurant (e.g. because of poor hygiene standards). Of course, this could also be explained in terms of the restaurant itself becoming associated with the occurrence of stomach ache, in a manner comparable to rats. Context-based renewal effects have been demonstrated in humans (e.g. Balooch & Neumann, 2011). At the time of writing this thesis, it was not clear how theory protection could be adequately tested against other explanations, in terms of context. Therefore, context is not investigated experimentally in the reported experiments.

Nevertheless, this does raise some interesting questions in terms of human learning and context. For example, would extinction occur more rapidly in a different context to excitatory training, compared to the same context (i.e. assuming that a different context could be used by participants to explain the absence of an outcome). There is also the question of whether a different pattern of results would be seen in rats.

In terms of existing models that have some similarity to theory protection, Pearce and Hall (1980) proposed an attentional model that does incorporate uncertainty (i.e. a lack of confidence), as a process that facilitates greater learning. Their model proposes that less learning occurs for cues that reliably predict outcomes. However, according to Pearce and Hall, learning is governed by the surprisingness of outcomes, while the concept of theory protection suggests that learning is governed by people's certainty about the causal status of cues. The Pearce and Hall model is considered in more detail later in this thesis, in relation to the experimental findings, and the difference between these approaches is unpacked further. Theory protection is also comparable to Bayesian models of associative learning (e.g. Courville, Daw, & Touretzky, 2006; Kruschke, 2008), in terms of being belief-focused. Such models represent beliefs about cues, in relation to differing hypotheses about what they might do. These beliefs are based on the likelihood of an outcome being caused by specific cues. Outcome likelihood varies according to experience (i.e. training), and Bayesian associative learning models weight values of likelihood according to the associative history of cues. However, rather than having a specific associative strength value, the status of cues is represented on a belief (i.e. probability) distribution, in which some associative strengths are more likely than others (e.g. it might be more likely that a cue causes stomach ache than not, or vice versa). One such Bayesian model, the Kalman filter (Kalman, 1960), has been

developed into a Bayesian equivalent of the Rescorla and Wagner (1972) model (e.g. see Kruschke, 2008), in which the predicted outcome is expressed as a distribution of beliefs across all possible outcome values. Similar to the Rescorla-Wagner model, this value is still based on an aggregate of the outcome predicted by all cues on a specific training trial, although this model is more flexible in its predictions, as it is not using fixed associative strength values. Confidence (or rather a lack of it) also plays a role in Bayesian inference, in that uncertainty about beliefs facilitates greater learning. Uncertainty is triggered by surprising outcomes (similar to Pearce and Hall, 1980). Again, this differs from theory protection, where a lack of confidence stems from the causal status of cues, on the basis of their ambiguity. It should be emphasised that Bayesian associative learning models are a broad approach, rather than a single specific theory that can be tested against theory protection. However, a Bayesian approach could be a route into formally implementing a theory protection-type process.

In the absence of a formal theory of causal confidence and associative learning, the starting point of the project reported in this thesis was the proposal of a simple theory protection account, for cases in which two cues are trained in compound. According to this account, human participants should engage in theory protection, by showing little or no learning about the cue that they are most confident about, while showing greater learning for the cue that they are least confident about (e.g. because its causal status is ambiguous in some way). The rationale for the experiments in Chapters 2-3 was the testable prediction that greater confidence about the causal status of cues, should lead to reduced learning. In other words, if participants are confident about the causal status of a cue, then they should protect this theory, rather than updating the causal status of that cue through subsequent learning. By extension, if participants are not confident about

the causal status of a cue, this lack of a strong theory should facilitate learning. It is possible to test this account against a simple prediction error account (i.e. individual prediction error, overall prediction error, or a mixture of both), by training two cues that are known to differ in their causal confidence (from the perspective of participants) in a compound, in order to compare the amount of learning for each cue.

All of the experiments reported in Chapters 2-3 were designed to test a prediction error account (e.g. Rescorla, 2001) against the theory protection account. These experiments utilised Rescorla's compound testing procedure (Rescorla, 2000; 2001) so that differences in learning about cues could be assessed. In each experiment, following an initial training stage, two cues were trained in compound, and the amount of learning for each cue was compared. The two cues always differed in terms of the size of their prediction error and how confident participants should be about their status. In all seven experiments, the expectation was that there would be less learning about the cue with the larger prediction error, since participants should protect their theory about that cue, on the basis of being more confident about its causal status. The experiments in Chapter 2 were based on a cue competition phenomenon called the redundancy effect (e.g. Uengoer, Lotz, & Pearce, 2013), while the experiments in Chapter 3 were based on the above Rescorla (2001) design. Furthermore, several of the Chapter 3 experiments were also designed to test attentional accounts, such as Mackintosh (1975) and Pearce and Hall (1980) against the theory protection account. In the final experiment, participants gave ratings of confidence about the causal status of cues, proving further support for the theory protection account. The experiments in Chapter 4 focused on a separate issue; uncertainty about the causal status of novel cues. The experimental data collected were used for a series of computational model fitting procedures, in order to find an adequate

formally implemented account of those data. However, the issues discussed in Chapter 4 link back to the issues discussed in Chapters 2-3. Chapter 5 considers the theoretical implications of all the research reported in this thesis.

2.1: Theory Protection 1

The experiments reported in Chapter 2 (and those reported in Chapter 4) are based on cue competition effects in associative learning. Such effects occur when the presence of one cue interferes with the learning about another cue. Blocking (Kamin, 1969) is the most widely known type of cue competition. It occurs when learning about a cue is apparently restricted by the simultaneous presence of another cue that has also been trained separately. For example, if a single cue is followed by an outcome (A+) , and a separately-encountered compound containing that cue is followed by the same outcome (AX+), then learning about X is restricted. In humans, learning is often tested by asking participants to rate the likelihood of the outcome, on the basis of specific cues (e.g. Jones, Zaksaite, & Mitchell, 2019). Participants rate blocked cues as a less likely cause of the outcome than an appropriate control (e.g. Y following B- BY+ training). There is more than one type of blocking. For example, in forward blocking, trials containing the separately trained cue are presented before the trails containing the compound (e.g. Shanks, 1985). Experiment 8 in Chapter 4 uses this kind of blocking. In simultaneous blocking, the trials containing the separately trained cue and those containing the compound are intermixed (e.g. Jones et al., 2019). Experiments 1 and 3 in this chapter use that kind of blocking. Overshadowing (Pavlov, 1927) is another type of cue competition. It is similar to blocking, in that learning is apparently restricted by the simultaneous presence of another cue, except that cue is not trained separately. For example, if a compound is followed by an outcome (AX+) but neither of those cues is encountered separately, then both cues will overshadow each other. Test ratings for both A and X should be lower than an appropriate control (e.g. K+), but also higher than for a blocked cue. Experiment 2 in this chapter uses overshadowing.

The aim of the first experiment reported in Chapter 2 was to pit the theory protection account (as proposed in this thesis) against a prediction error account (e.g. Rescorla, 2001). To achieve this, cues were chosen that had specific properties. Firstly, one cue needed to have a more ambiguous causal status than the other, so that the theory protection account would predict greater learning for the more ambiguous cue. Secondly, the cue with the least ambiguous causal status needed to be judged as a less likely cause of the outcome prior to the compound conditioning phase. This was to ensure that a prediction error account would predict greater learning for this cue, when the compound was trained as a cause of the outcome. Two cues with these properties are found in the redundancy effect (e.g. Uengoer et al., 2013).

A typical redundancy effect design involves presenting participants with a training stage incorporating blocking (A+ AX+) and a simple discrimination (BY+ CY-). Cue Y, from the simple discrimination, is referred to as an uncorrelated cue, because it appears in both a causal compound and a non-causal compound. The redundancy effect is the robust finding that X is rated as a more likely cause of the outcome than Y during a subsequent test phase (Jones & Zaksaite, 2018; Jones et al., 2019; Uengoer, Dwyer, Koenig, & Pearce, 2019; for analogous results in non-human animals, see Jones & Pearce, 2015; Pearce, Dopson, Haselgrove, & Esber, 2012). Crucially, there is evidence from Jones et al. (2019) that participants' confidence about the causal status of X and Y differs at test. Participants were asked to make confidence ratings during the test stage of a redundancy effect experiment, in addition to outcome likelihood ratings. After participants had rated the likelihood of the outcome for specific cues, they were asked to rate how confident they were about their likelihood ratings. The mean confidence

ratings for the blocked cue (X) were significantly lower than those for the uncorrelated cue (Y), suggesting that participants were less confident about the causal status of X than Y. A further experiment by Jones et al. (2019) showed that the likelihood ratings of blocked cues, but not uncorrelated cues, were dependent on the overall proportion of training trials on which the outcome occurred. The results showed higher ratings when the proportion of trials on which the outcome occurred was higher. This suggests that participants' beliefs about X were more labile than for Y, further supporting the idea that participants are less confident about the causal status of blocked cues than uncorrelated cues.

It is worth noting that this theoretical view, in which participants are unconfident (i.e. uncertain) about the causal status of blocked cues, is established within the learning literature. For example, uncertainty about the causal status of blocked cues is supported by several studies investigating the effects of manipulating assumptions about the outcome (e.g. Lovibond, Been, Mitchell, Bouton, & Frohardt 2003; Beckers, De Houwer, Pineno, & Miller, 2005; Vandorpe, De Houwer, & Beckers, 2007). In terms of the likelihood ratings provided by participants, cue X is typically assigned ratings in the middle of an 11-point likelihood scale, while Y is given lower likelihood ratings, suggesting that participants learn that Y is unlikely to be a cause of the outcome. On the basis of the aforementioned confidence data, the intermediate likelihood ratings given to X support the view that participants are unconfident about the status of X, despite encountering it in a causal compound (AX+). In other words, participants might give such intermediate likelihood ratings when they are uncertain because they lack the confidence to assign either a high or a low rating. Nevertheless, this difference in likelihood ratings suggests that, if X and Y were combined in a subsequent XY+

training phase, greater learning for Y than for X would be consistent with a prediction error account.

Three experiments are reported in this chapter. In each one, following an initial training stage, a blocked (or overshadowed) cue was trained in compound with an uncorrelated (or negative discriminator) cue. The compound was always followed by presence of the outcome. The cue with the larger prediction error at the start of the compound training phase was always the cue about which participants should hold a stronger theory. This allowed a direct comparison of the theory protection account against a prediction error account, since these theories make opposing predictions in each of the reported experiments.

2.2: Experiment 1

As with previous redundancy effect studies using human participants (e.g. Uengoer et al., 2013), this experiment used a food allergy scenario. Participants were required to learn whether an allergic reaction would occur, on the basis of different single foods or pairs of foods being eaten. They were presented with a fictional scenario, in which they played the role of a medical doctor, trying to ascertain which foods cause a stomach ache in a test patient. During the training trials, participants were presented with a series of food cues and were asked to predict whether or not the fictional patient would experience a stomach ache after eating these foods. After participants made their prediction, they were then provided with feedback as to whether or not a stomach ache occurred. Following training, participants were tested by being asked to make a likelihood rating indicating how likely they thought a stomach ache would be after the patient ate specific foods.

The design of the experiment is shown in Table 1. This design contained two blocked cues (W and X) and two uncorrelated cues (Y and Z). Following this training phase, one blocked cue and one uncorrelated cue were trained together and paired with the outcome (XY+). If learning in this phase is the result of individual prediction error, in a similar manner to Rescorla (2001), there should be more learning about Y than X as it will have the greater error at the start of the XY+ stage. This is because uncorrelated cues are given lower likelihood ratings than blocked cues, so this lower rating is the most discrepant with the outcome when XY+ is presented. Alternatively, if learning is determined by overall prediction error, there should be equal learning about X and Y. Finally, if learning is determined by the theory protection account, then participants should protect their theory about Y since they have already learned that it is not a cause

of the outcome with relative confidence (Jones et al., 2019). However, they should readily attribute the outcome to the causally ambiguous X during XY+ trials. Consequently, X should be learned about more than Y.

Table 1. The design of Experiments 1-3

Experiment	Stage 1	Stage 2	Test
1	A+ AX+ BY+ CY-	XY+	XZ WY
	D+ DW+ EZ+ FZ-		A B C D E F X Y W Z
2	AX+ BY+ CY-	XY+	XZ WY
	DW+ EZ+ FZ-		A B C D E F X Y W Z
3	A+ AX+ BY+ CY-	XC+	XF WC
	D+ DW+ EZ+ FZ-		A B C D E F X Y W Z

Key:
Letters A-F = different cues
+ = stomach ache
- = no stomach ache

Learning about X and Y was compared using a final discrimination of a similar kind to that used by Rescorla (2001). As well as being asked for likelihood ratings for each individual cue, participants were asked to rate the likelihood of the outcome for two compounds, XZ and WY. Each of these compounds contained one cue that had been blocked in Stage 1, and one that had been an uncorrelated cue in Stage 1. In the absence of Stage 2 training, these two compounds should be assigned the same likelihood ratings at test. Consequently, any difference between these compounds must necessarily be the result of the XY+ training in Stage 2, and would indicate a different amount of learning about X and Y during that stage. Therefore, if learning is governed by an individual prediction error term, participants should rate the likelihood of the outcome as being higher for WY than XZ. If learning is governed by an overall error term, then there should either be no difference between ratings of XZ and WY, or a higher rating

for WY if a modified response function (i.e. Holmes et al., 2019) is assumed. However, if learning is determined by the theory protection account, then XZ should be assigned higher ratings than WY at test.

2.2.2: Method

Participants

Thirty-six psychology students from the University of Plymouth participated in this experiment, in return for course credit (30 female, 6 male; mean age = 20.2, SD = 3.1). This sample size has adequate power to detect medium-sized within-subjects effects (83% power at d = 0.5). People who had previously taken part in similar experiments were excluded from this study, to ensure participants were naive to the purpose of the experiment.

Materials

Participants were all tested in the same lab at University of Plymouth. The experiment was conducted using Viglen Genie desktop computers, running the Windows 10 operating system. The computers all used 22-inch Phillips LED displays, with participants at a typical distance (of approximately 40-80 cm) from the screen. The experiment was designed and executed in the Psychopy desktop application version 1.83.04 (Peirce, 2007), with the output generated as individual CSV files for each participant. Participants made their responses by pressing keys on a standard UK computer keyboard during the training stages, and by using mouse clicks during the test stage. The ten individual cue types were represented on screen as photographs of fruits: apple, banana, cherry, kiwi, mango, orange, peach, pear, plum and strawberry. All the fruits were presented within a white square. The dimensions of each cue (including the white square) were 300 x 300 pixels, with a screen resolution of 1920 x 1080 pixels. For each participant, the foods were randomly assigned to each cue (A, B, C, D, E, F, X, Y,

W and Z). The two outcomes, ‘stomach ache’ and ‘no stomach ache’, were represented by text on screen and a photograph of a man clutching his stomach, or a man giving a ‘thumbs up’, respectively. The outcome images were presented within a white rectangle. The dimensions of the outcome images (including the white rectangle) were 291 x 332 pixels. All experimental text, including instructions, was white. A black background was used throughout the experiment. Study information sheets, consent forms and debrief forms were all printed on paper.

Design

The experiment used a within-subjects design, as outlined in Table 1. During Stage 1, participants were presented with twelve blocks of training. The eight trial types (A+, AX+, BY+, CY-, D+, DW+, EZ+, FZ-) appeared in a random order within each block. Each trial type was only presented once within each block. There were six trials in Stage 2, all with XY+. During the Test stage, participants were presented with two blocks of test cues. The twelve trial types (A, B, C, D, E, F, X, Y, W, Z, XZ and WY) appeared in a random order within each block. Each trial type was only presented once within each block.

Procedure

Participants were required to read an information sheet and sign a consent form prior to participating in the experiment. The experimental instructions were presented on the screen at the start of the experiment. They were adapted from Uengoer et al. (2013) and were as follows:

This study is concerned with the way in which people learn about relationships between events. In the present case, you should learn whether the consumption of certain foods leads to stomach ache or not.

Imagine that you are a medical doctor. One of your patients often suffers from a stomach ache after eating. To identify which foods they react to, the patient eats specific foods and observes whether a stomach ache occurs or not. The results of these tests are shown to you on the screen one after the other.

You will always be told what your patient has eaten. Sometimes, they have only consumed a single kind of food and on other times they have consumed two different foods. Please look at the foods carefully.

You will then be asked to predict whether the patient suffers from stomach ache. For this prediction, please click on the appropriate response button. After you have made your prediction, you will be informed whether your patient actually suffered from stomach ache. Use this feedback to find out what foods cause a stomach ache in your patient. At first you will have to guess the outcome because you do not know anything about your patient. But eventually you will learn which foods lead to stomach ache in this patient and you will be able to make correct predictions.

For all of your answers, accuracy rather than speed is essential. Please do not take any notes during the experiment. If you have any more questions, please ask them now. If you do not have any questions, please start the experiment by pressing the space bar.

For each trial during the training stages, the cues were presented visually on either the left- or right-hand side of the screen. When only one image was presented, the opposite side of the screen contained a blank space. The cues were randomly assigned to either the left or right position on each trial. Text at the top of the screen stated that ‘The patient eats the following:’, with the stimuli presented below this. Underneath the stimuli, further text stated ‘Which outcome do you expect? Please use your keyboard to respond’. Participants were instructed to respond by pressing the appropriate key on their keyboard; Z for ‘No Stomach Ache’ and M for ‘Stomach Ache’. After participants made their response, the feedback for that trial was shown. The feedback screen consisted of the appropriate outcome image along with its accompanying text, indicating either ‘Stomach Ache’ or ‘No Stomach Ache’. The feedback was shown on screen for two seconds, after which the next trial began.

After the completion of Stage 1, Stage 2 started with no trial break so that from the perspective of participants this was a seamless continuation of the training. This stage consisted of a previously unseen compound XY presented six times in a row. As in Stage 1, the cues were randomised on each trial to appear on either the left- or right-hand side of the screen. The on-screen text and responding via the keyboard was the same as in Stage 1. The process for displaying the trial feedback was also the same, except that all six trials resulted in ‘Stomach Ache’ as the outcome.

After the completion of Stage 2, a further instruction screen was shown before commencement of the Test stage:

Next, your task is to judge the probability with which specific foods cause stomach ache in your patient. Single foods and pairs of foods will be shown to you on the screen. In this part of the experiment, you will receive no feedback about the actual reaction of the patient. Use the information that you have collected so far, to make your rating. Press the space bar to continue the experiment.

For each trial during the Test stage, the cues were presented on either the left- or right-hand side of the screen. When only one image was presented, the opposite side of the screen again contained a blank space. The cues were randomly assigned to either the left or right position on each trial. As before, text at the top of the screen stated that 'The patient eats the following:', with the stimuli presented below this. Underneath the stimuli, further text stated 'How likely are they to suffer a stomach ache? (0 = Very Unlikely; 10 = Very Likely)'. Participants were instructed to respond by clicking on an eleven-point rating scale using their mouse pointer, to indicate how likely they thought the occurrence of a stomach ache would be. The rating scale was located in the lower part of the screen, with the 11-point scale running from left to right, in ascending numerical order. After participants made their response, a black screen appeared for 0.4 secs, after which the next trial was presented. Following the completion of the experiment, participants were provided with a debrief form.

Analysis

The data were processed and analysed using R (R Core Team, 2018). The difference between the XZ and WY test compounds was assessed using paired-samples t-tests. Some additional analyses were also conducted on key single test stimuli, to test for

specific predicted differences between cue ratings. The alpha level was set to $p < .05$ for all tests. As these tests were done on the basis of specific prior predictions, there was no requirement for Bonferroni corrections. Bayesian t-tests were also conducted, using the procedure recommended by Dienes (2011) and implemented as R code by Baguley and Kaye (2010). As there was no suitable previous study on which to specify a plausible predicted effect size, a uniform distribution was specified, with a lower limit of -5 and an upper limit of 5 (in terms of the mean difference between ratings). The motivation for this was that previous human experiments utilising this compound testing procedure (e.g. Mitchell, Harris, Westbrook, & Griffiths, 2008) produced mean differences in compounds lower than 5 (where 5 would be considered a reasonable limit to any observed difference when an 11-point rating scale is used). For Bayesian t-tests conducted on single cues, the lower limit was set to -10 and the upper limit was set to 10, since these are the largest mean differences in either direction permitted by an 11-point scale. In keeping with accepted conventions (e.g. Jeffreys, 1961), a Bayes factor of over three was set as the level providing evidence for a difference, while a Bayes factor of less than one third was set as the level providing evidence for no difference. Values between these levels were accepted as being inconclusive.

2.2.3: Results and Discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/4xbkp/. The descriptive statistics for the Experiment 1 training stages are shown in Figure 1. The Stage 1 data indicate that participants learned sufficiently about the eight different trial types by the time first training stage was complete. Similarly, the Stage 2 data indicate that participants learned that the XY+ compound was causal by the end of the second training stage, after giving it an intermediate rating on the first trial.

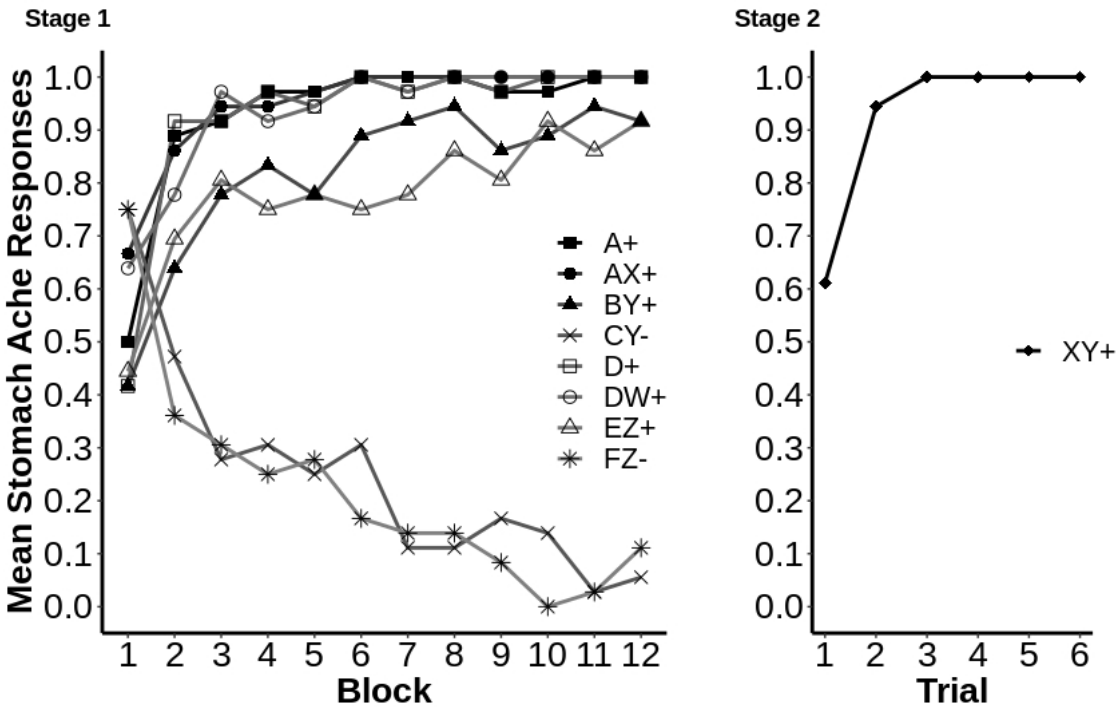

Figure 1. Experiment 1 Stage 1 and 2 data.

Descriptive statistics for the Experiment 1 Test stage are shown in Figure 2. Ratings for XZ were significantly higher than for WY; $t(35) = 2.99$, $p = .005$, $BF = 15.62$, $d = .50$. Further testing revealed higher ratings for the single cue X compared to W; $t(35) = 3.96$, $p < .001$, $BF = 215.66$, $d = .66$. Conversely, there was no difference between the ratings assigned to Y and Z; $t(35) = 0.53$, $p = .597$, $BF = .08$, $d = .09$, and therefore no evidence

that participants learned about Y during the second stage of the experiment. Taken as a whole, these findings show that the differences between the compounds were specifically driven by learning about X during XY+ trials. Figure 2 panel B shows inter-subject variability on the key compound test difference. As would be expected from the mean test ratings, most individual participants rated XZ higher than WY, although the variability did extend to some participants assigning a higher rating to WY. These data are consistent with the theory protection account, as opposed to a prediction error account (Rescorla, 2001). Despite the theoretical implications of this result, the crucial next step was to ensure its generality. Experiment 2 was intended as an extension of Experiment 1, to increase the generality of the findings. The design of Experiment 2 was modified, to vary the type of causally ambiguous cue incorporated into the design.

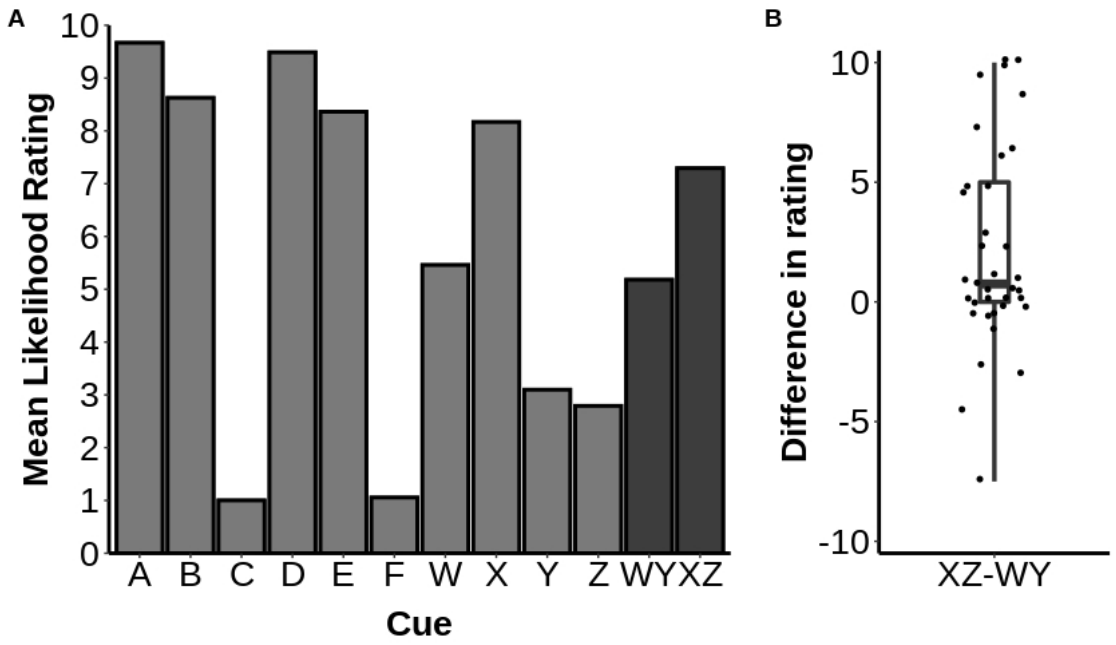

Figure 2. Experiment 1 Test stage ratings for all single stimuli and the two compound cues. Panel B shows inter-subject variability on the key XZ-WY difference using a method comparable to Hintze & Nelson (1998). Each dot is one participant, with jitter applied for readability. The boxplot shows the median and interquartile range.

2.3: Experiment 2

There were two key aims for Experiment 2. The first was to extend the generality of the findings of Experiment 1 by using a different kind of causally ambiguous cue. To achieve this, the experimental design was modified so that it incorporated an overshadowed cue (cf. Waldmann, 2001) from a two-item compound, rather than a blocked cue. As stated in the chapter introduction (2.1), overshadowed cues are similar to blocked cues, in that two cues are trained in compound, but neither cue is presented separately. The logic was that, like blocked cues, overshadowed cues are causally ambiguous from the perspective of the participant. This is supported by Jones et al. (2019), who reported that participants were less certain about their causal judgements of overshadowed cues than of uncorrelated cues. Presumably this is because, when two cues are presented in compound and the outcome is present (e.g. AX+, without A+ training), participants do not know which of the two cues is the cause of the outcome. However, unlike blocked cues, compound trials containing a pair of overshadowed cues do at least allow participants to infer that at least one, or both, of the cues must be a cause of the outcome. Accordingly, Jones et al. observed higher causal ratings for overshadowed cues than for blocked cues. Learning during Stage 2 should again be greater for X (the overshadowed cue) rather than Y, because of the difference in confidence participants should have, in their theory about these two cues. However, the higher causal ratings given to overshadowed cues, compared to blocked cues, meant that there was a theoretical basis for expecting less learning during the XY+ stage, since there was less discrepancy to be accounted for.

The design of the experiment is shown in Table 1. All details were identical to Experiment 1, except that the A+ and D+ trials were omitted from Stage 1. Following

Experiment 1, the expectation was that causal ratings during the Test stage would again be higher for XZ than for WY. However, on the basis of the higher causal ratings reported by Jones et al. (2019) for overshadowed cues than for blocked cues, the expectation was that the size of this effect would be somewhat smaller than in Experiment 1.

2.3.2: Method

Participants

Forty participants were recruited from the University of Plymouth, in return for a small monetary payment (32 female, 8 male; mean age = 26.7, SD = 10.4). This sample size has adequate power to detect medium-sized within-subjects effects (87% power at d = 0.5). People who had previously taken part in similar experiments were excluded from this study, to ensure participants were naive to the purpose of the experiment.

Materials

The materials used for Experiment 2 were the same as those used for Experiment 1, except that Psychopy version 1.85.1 was used (Peirce, 2007).

Design

The experiment used a within-subjects design, as outlined in Table 1. During Stage 1, participants were presented with twelve blocks of training. The six trial types (AX+, BY+, CY-, DW+, EZ+, FZ-) appeared in a random order within each block. Each trial type was only presented once within each block. Stage 2 and the Test stage were exactly the same as in Experiment 1.

Procedure and analysis

The procedure for Experiment 2 was the same as for Experiment 1. The analyses were the same, except that the Bayesian priors were updated on the basis of the results of Experiment 1. As outlined above, there was a theoretical basis for expecting the Experiment 2 mean differences to be smaller than the values observed in Experiment 1 (because of the higher ratings given to overshadowed cues than blocked cues). Therefore, following the recommendations of Dienes (2011), a half-normal prior distribution was specified for each Bayesian t-test, with a mean of zero and a standard deviation set to the mean differences observed in Experiment 1.

2.3.3: Results and discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/8ceub/. The descriptive statistics for the Experiment 2 training stages are shown in Figure 3. The Stage 1 data indicate that participants learned sufficiently about the six different trial types by the time first training stage was complete. Similarly, the Stage 2 data indicate that participants learned that the XY+ compound was causal by the end of the second training stage.

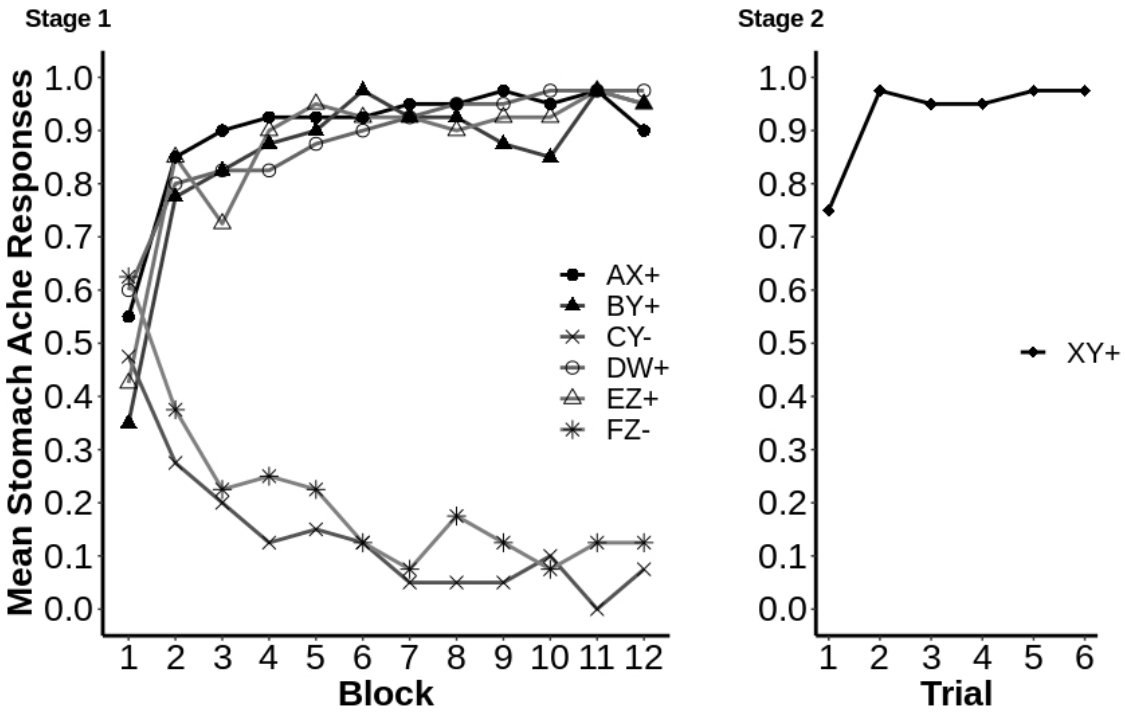

Figure 3. Experiment 2 Stage 1 and 2 training data.

The descriptive statistics for the Test stage are displayed in Figure 4. Ratings for XZ were significantly higher than for WY; $t(39) = 2.37$, $p = .023$, $BF = 6.42$, $d = .37$. This finding supports the prediction that the compound containing X would receive higher ratings. The difference between the compounds was again reflected in higher ratings for X compared to W; $t(39) = 2.37$, $p = .023$, $BF = 4.6$, $d = .38$. There was no evidence for a

significant difference between Y and Z, despite the ratings for Y appearing slightly lower; $t(39) = 1.57$, $p = .124$, $BF = 1.13$, $d = .25$. As with Experiment 1, this finding supports the theory protection account, rather than a prediction error account (Rescorla, 2001). Although the ratings for W appeared to be slightly higher than for A or D, there was no evidence for a significant difference between either A and W, $t(39) = 1.17$, $p = .249$, $BF = .58$, $d = .19$; or D and W, $t(39) = 1.72$, $p = .093$, $BF = 1.09$, $d = .27$.

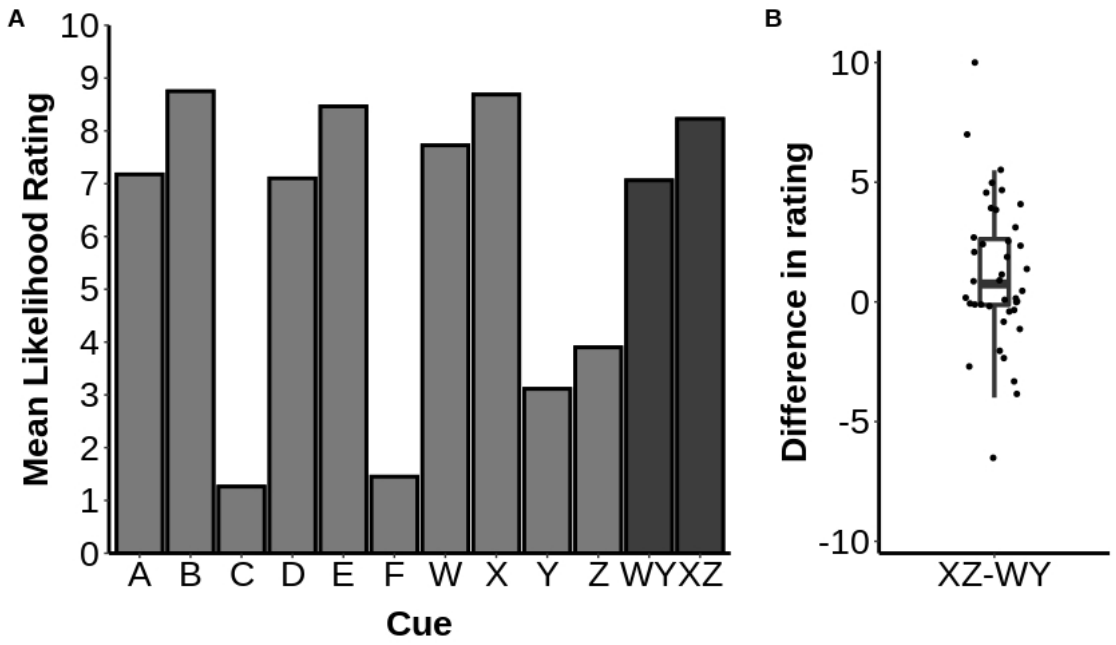

Figure 4. Experiment 2 Test stage ratings for all single stimuli and the two compound cues. Panel B is a plot showing inter-subject variability on the key XZ-WY difference.

The results of Experiment 2 are consistent with those from Experiment 1; in both cases, participants learned more about the cue about which they should have held a weaker theory, as opposed to the cue that should have had the greater prediction error, when the two cues were trained in compound. However, one limitation of these experiments is that the number of trials featuring the two critical cues during Stage 1 was not matched. Cue Y was presented to participants twice as often as X, because it was included in both BY+ and CY- trials. This might be important because, according to some theories of

attention (e.g. Mackintosh, 1975; Pearce & Hall, 1980), learning rate is influenced by the amount of prior exposure to each cue. Another limitation is that the causal status of Y could be considered somewhat ambiguous, since its occurrence during Stage 1 was followed equally often by stomach ache and no stomach ache. Accordingly, Rescorla and Wagner's (1972) model predicts that Y will maintain some associative strength during the first stage of Experiment 1. It does so because C is predicted to become an inhibitor for the outcome, protecting Y from extinction on CY- trials. In fact, Rescorla and Wagner's model predicts that Y should have more associative strength than X in many circumstances. Although the opposite result has been observed numerous times (e.g. Uengoer et al., 2013), this is a reason to treat assumptions about the causal status of Y with caution. In Experiment 3 these issues were addressed by comparing learning about a blocked cue with learning about a different cue (about which participants would be expected to hold a confident theory); one that was presented the same number of times as the blocked cue, and that was never presented with the outcome during Stage 1.

2.4: Experiment 3

The design of Experiment 3 is shown in Table 1. Having extended the generality of the findings of Experiment 1 by comparing learning about Y to a different causally ambiguous cue in Experiment 2, the next objective was to check that the effect would persist if a different causally non-ambiguous cue was used in Stage 2. Training during Stage 1 was the same as for Experiment 1, but Stage 2 differed in Experiment 3, in that participants received XC+ training. The inclusion of C in this compound was motivated by the high confidence ratings observed by Jones et al. (2019) for this cue compared to X, indicating that participants have a stronger theory about the causal status of C. It was also motivated by an assumption that the prediction error for C on the CX+ trials would be large. Evidence for the latter assumption is provided by the consistently low causal ratings assigned to C in previous experiments (Experiments 1 and 2, and all experiments reported by Jones et al., 2019; and Uengoer et al., 2013), and the fact that C was always presented without the outcome during Stage 1. The Test stage contained two compounds, XF and WC, that permitted a comparison of learning about X and C according to the same logic as the previous experiments. If learning is determined by individual prediction error, participants should learn more about C than X during Stage 2, and causal ratings should be higher for the WC compound than for the XF compound during the Test stage (or no difference for overall prediction error). Alternatively, if participants in Experiments 1 and 2 learned most about X during Stage 2 because of their lack of a theory about its causal status, then that effect should persist in Experiment 3, leading to higher causal ratings for XF than for WC at test.

2.4.2: Method

Participants

Forty Psychology students from the University of Plymouth participated in this experiment, in return for course credit (37 female, 3 male; mean age = 22.6, SD = 7.0). This sample size has adequate power to detect medium-sized within-subjects effects (87% power at d = 0.5). People who had previously taken part in similar experiments were excluded from this study, to ensure participants were naive to the purpose of the experiment.

Materials

The materials used for Experiment 3 were the same as those used for Experiments 1 and 2, except that Psychopy version 1.85.2 was used (Peirce, 2007).

Design

The experiment used a within-subjects design, as outlined in Table 1. During Stage 1, participants were presented with twelve blocks of training. The eight trial types (A+, AX+, BY+, CY-, D+, DW+, EZ+, FZ-) appeared in a random order within each block. Each trial type was only presented once within each block. There were six trials in Stage 2, all with XC+. During the Test stage, participants were presented with two blocks of test cues. The twelve trial types (A, B, C, D, E, F, X, Y, W, Z, XF and WC) appeared in a random order within each block. Each trial type was only presented once within each block.

Procedure and analysis

The procedure for Experiment 3 was the same as for Experiments 1 and 2. The analyses were the same, except that the Bayesian priors were updated on the basis of the results of Experiment 1. There was a theoretical basis for expecting the Experiment 3 mean differences to be about the same as the values observed in Experiment 1, in that the abstract design of the first training stage was the same. Therefore, following Dienes (2011), a normal distribution was specified as the prior for each Bayesian t-test, with each mean set to the corresponding Experiment 1 mean and each standard deviation set to half this value.

2.4.3: Results and discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/jqrb6/. The descriptive statistics for the Experiment 3 training stages are shown in Figure 5. The data from Stage 1 indicate that participants learned sufficiently about the eight different trial types by the time training was complete. Similarly, the data from Stage 2 shows that participants learned that the XC+ compound was causal by the end of that stage, after giving it an intermediate rating on the first trial.

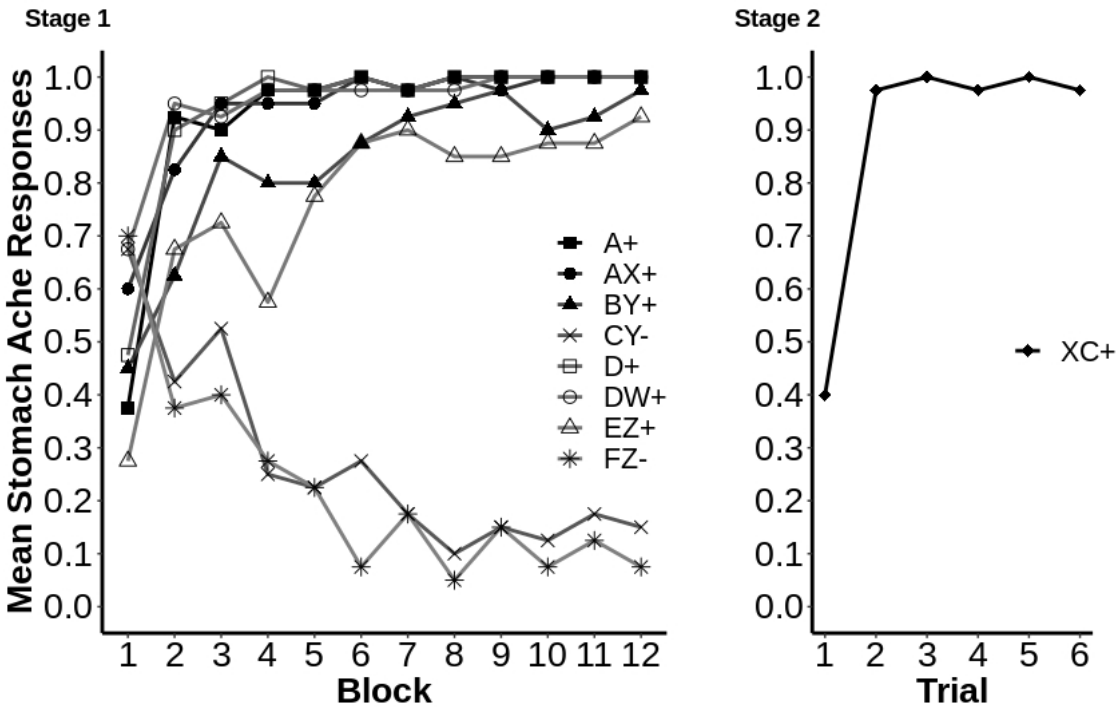

Figure 5. Experiment 3 Stage 1 and 2 data.

Descriptive statistics for the Test stage are displayed in Figure 6. Causal ratings were significantly higher for XF than for WC; $t(39) = 2.39$, $p = .022$, $BF = 6.66$, $d = .38$. This finding supports the prediction that the compound containing X would be rated higher at test. As expected, the difference between the compounds was driven by higher ratings for X compared to W; $t(39) = 3.42$, $p = .001$, $BF = 128.57$, $d = .54$. Although the ratings

for C looked a little higher than for F, there was no evidence for a significant difference; $t(39) = 1.93$, $p = .061$, $BF = .70$, $d = .31$. These results are consistent with theory protection, with participants learning more about the causally ambiguous blocked cue (X) than the causally non-ambiguous discriminative cue (C). As with the previous experiments, this finding is incompatible with a prediction error account (Rescorla, 2001).

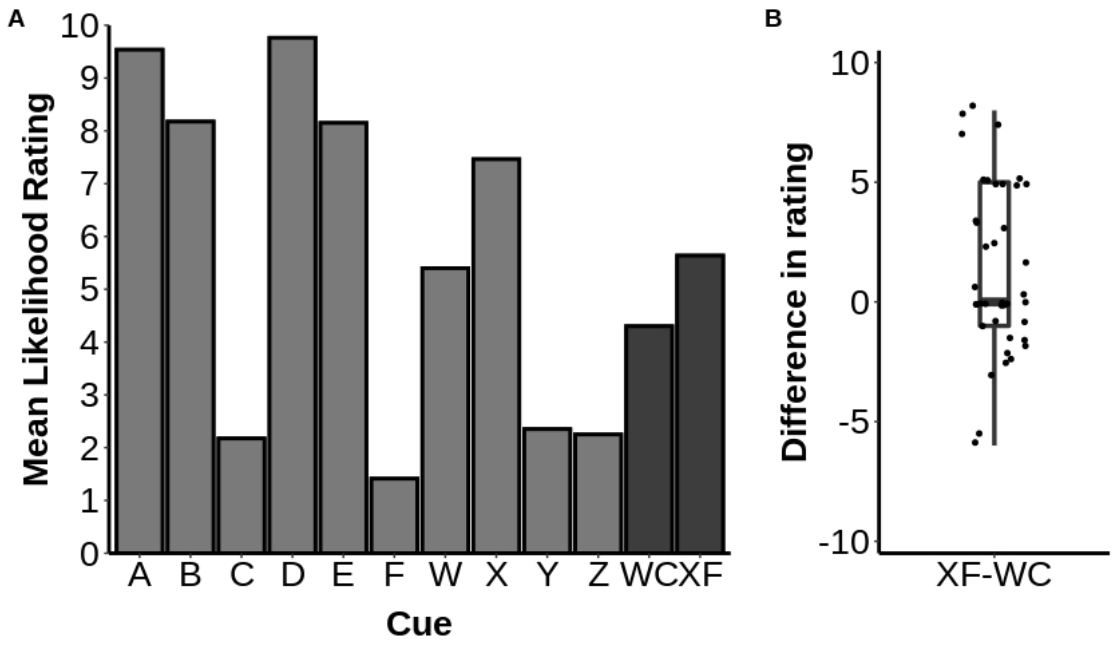

Figure 6. Experiment 3 Test stage ratings for all single stimuli and the two compound cues. Panel B is a plot showing inter-subject variability on the key XF-WC difference.

2.5: General Discussion

The three experiments in this chapter provide evidence that causal learning in humans is, at least in part, governed by theory protection. If participants hold a stronger theory about a cue, then they appear to protect this theory to a greater extent, compared to cues about which they do not hold a strong theory. Therefore, differing confidence about the causal status of cues appears to govern which cues are learned about the most, rather than differences in the size of the prediction error. In each experiment, participants apparently resisted updating their beliefs about cues with a known causal status, instead attributing unexpected outcomes to cues with a comparatively ambiguous causal status. This occurred despite the ambiguous cues having the smaller prediction error at the start of the second training stage in all three experiments. The results of these experiments are the opposite to those that might be expected on the basis of prediction error accounts of learning (Bush & Mosteller, 1951; Rescorla & Wagner, 1972; Rescorla, 2001). A theoretical implication of particular interest is the idea that humans act to protect their theories about known causal relationships. This idea is intuitive, in terms of learning being a process of acquiring information about the things we are unsure about and incorporating this information alongside existing knowledge.

The theory protection account provides an alternative explanation for previous experiments using compound testing procedures in humans. For example, Mitchell et al. (2008) initially gave participants A+ C+ training, and then subsequently trained cue A and a novel cue B in a causal compound (AB+). A forced choice test on two compounds (AD versus BC) revealed evidence of more learning about the novel cue B during AB+ trials. This effect was subsequently replicated with a larger outcome (AB++) in the second training stage, again showing more learning about B than about A. Although

these data are consistent with a prediction error account, they are also consistent with the theory protection account, showing that these accounts need not make opposing predictions in all cases. This is because B would have both the greater prediction error and the more ambiguous causal status at the start of the second training stage, as nothing had yet been learned about it. In another experiment, Le Pelley and McLaren (2001) initially trained two cues as excitors (A+ C+), and two other cues (B and D) as inhibitors (BE- DE- E+). One cue of each type was used in the second training phase (AB+); the results of a subsequent compound test revealed that participants had learned more about cue A than cue B, despite the latter presumably having a greater prediction error. Although participants should have been confident about the causal status of both A and B following initial training, these findings are still compatible with the theory protection account because participants appeared to maintain (and strengthen) their belief that A was a cause of the outcome but B was not. This suggests that there may be more to theory protection than just differences in confidence governing learning; the discussion of Chapter 3 (3.6) and parts of Chapter 5 consider the broader manner in which theory protection might operate, and the role for causal confidence within such a framework.

While the results of the present experiments appear to be inconsistent with a prediction error account, they might be accommodated if an additional process is invoked. One obvious candidate is the modification of attention to cues as a result of experience. For example, Mackintosh's (1975) model of learning is compatible with the results of Experiments 1 and 2. According to Mackintosh's model, more attention will be paid to cues that are better predictors of outcomes rather than poorer predictors, resulting in more learning about these cues. Rather than this attentional process being an alternative

to prediction error, it is instead suggested to operate alongside individual prediction error, with attention determining associability. It is not necessary to know the equation for this model, for the purposes of this thesis, but it is included in an addendum at the end of Chapter 5. In Experiments 1 and 2, X might be considered more predictive of the outcome than Y during Stage 1. This is because X was consistently followed by the outcome, whereas Y was followed by the outcome on BY+ trials but not CY- trials. If participants paid more attention to X than to Y as a result, this could have led to greater learning about X than Y during Stage 2. Experiment 3, however, is harder to reconcile with Mackintosh's model. Cue C was a better predictor of the absence of the outcome than Y on CY- trials, so the model predicts that participants should have learned to pay attention to C during Stage 1. Although X was consistently paired with the outcome on AX+ trials, it was a poorer predictor of that outcome than A (because of the separate A+ trials) and should therefore have suffered at least some decline in attention during Stage 1. Consequently, greater learning for X than for C during Stage 2 cannot have been the result of changes in attention that occurred during Stage 1. Mackintosh's model can only be reconciled with the results of Experiment 3 if one assumes that the relevant changes in attention occurred during Stage 2. Before the first XC+ trial, C had only been paired with the absence of the outcome. When it was subsequently presented in the XC+ compound, it was therefore a poor predictor of the outcome. This could have resulted in a rapid decline in the associability of C relative to X, leading to more learning about X than C from the second trial onwards, in spite of C having the larger prediction error. However, this account is not readily supported by the XC+ training data (see Figure 5). If these data are taken at face value, almost all the learning about the XC+ compound appears to have taken place during the first Stage 2 trial, before any update in associability could have influenced learning. This interpretation should be treated with

caution, however, since high causal ratings for XC+ from the second trial onwards do not necessarily imply that learning was complete. To investigate this point further, an obvious next step would be to conduct a similar experiment to Experiment 3, but with only one XC+ training trial in Stage 2. Two pilot experiments have been conducted using this procedure but the results were inconclusive. The data and analyses for these two unpublished experiments are available online at https://osf.io/8f3a7/ and https://osf.io/9dvcf/ Mackintosh's model is instead tested against the theory protection account in Chapter 3, using a different approach.

An additional way in which Mackintosh's (1975) theory could be reconciled with the results in this chapter, is if one assumes that cues given high and low causal ratings differ in their associability changes. For example, it may be that cues given low causal ratings acquire associability at a slower rate than those given high causal ratings. If this were the case, then it would be possible for X to have higher associability than C at the start of Stage 2. The most obvious way to test this would be to train participants with one cue that is a cause of an outcome, which also has a non-ambiguous causal status, alongside another cue that does not cause that outcome, but which is less predictive and has a more ambiguous status. If both cues were then trained in a compound that does not cause the outcome in Stage 2, Mackintosh's model would predict a more substantial decrease in associative strength for the previously predictive causal cue. The theory protection account, on the other hand, would predict more learning about the causally ambiguous cue. This idea is also tested directly in Chapter 3.

An alternative view of how attention changes as a result of experience was provided by Pearce and Hall (1980). They proposed that animals pay attention to cues that are

followed by surprising outcomes, and that cues that are followed by predicted outcomes suffer a decline in attention. The associability is calculated by subtracting the sum of the associative strengths on the previous trial from the asymptote of learning. It is not necessary to know the equation for this model either, but it is also included in the addendum at the end of Chapter 5. As stated in Chapter 1, Pearce and Hall's model does bear some conceptual resemblance to the theory protection account. Both accounts suggest that there should be less learning about cues that have known consequences, encapsulating the idea that learning is a process of reducing uncertainty about our environment. However, the mechanisms are quite different. In the case of Pearce and Hall's model this process is outcome-directed, as it is the surprisingness of the outcome that governs any update in associability (and consequently associative strength) for all cues that are present. In the case of the theory protection account, it is knowledge about the status of cues themselves, rather than outcomes, that influences future learning. In other words, learning is driven by knowledge about individual cues, rather than being driven by knowledge about the outcomes that simultaneously presented cues are paired with. This difference in focus means that the two theories make differing predictions for the present experiments. For example, by the end of Stage 1 of Experiment 3, Pearce and Hall's model predicts a decline in attention for all cues because the outcome is predictable on every trial. The training data from this experiment suggest that, if anything, the presence or absence of the outcome was better predicted on trials containing X than trials containing C. As a result, the associability of each cue should have been low at the outset of Stage 2, with a possible advantage for C. This is inconsistent with the observation of greater learning for X than for C in Stage 2. The theory protection account predicts more learning about X than about C because the causal status of X was ambiguous at the end of Stage 1, even though the outcome was

predictable on AX+ trials. Pearce and Hall's model similarly predicts that all cues in Stage 1 of Experiments 1 and 2 should decline in associability. This would lead to equal learning about both cues in Stage 2. The Mackintosh, and Pearce and Hall attentional models are investigated further in Chapter 3.

It is worth briefly mentioning that, somewhat unusually for studies of predictive learning, plots of inter-subject variability were provided for the key tests. These plots suggest that a substantial minority of participants may be behaving in a way consistent with prediction error accounts (i.e. they individually have an XZ-WY difference that is zero or negative). The amount of data collected per participant in the current experiments precludes any examination of whether this is a stable individual difference; it could, alternatively, be measurement error. Future research might examine this issue of the presence of stable individual differences by lengthening test stages, or testing the same people across multiple procedures.

3.1: Theory Protection 2

According to the theory protection account, learning is influenced by the extent to which participants already have a theory about what cues do. Instead of learning according to how large each prediction error is, human participants should maintain existing causal associations, as far as is possible, and attribute unexpected outcomes to cues about which they are not confident. To return to the example of medication from Chapter 1, if you have already learned that a particular medication causes a headache (A+) then you might be resistant to updating your beliefs when further consumption of A does not result in a headache. If you have taken this medication in combination with something else (AB-), you should attribute the absence of the headache more readily to the added substance B, provided that this attribution does not contradict an existing theory about B.

As already stated, the theory protection account has obvious implications in human learning for the Rescorla (2001) experiment described in Chapter 1. Recall that, following A+ (and C+) training in the first stage, a compound of cue A and previously non-reinforced cue B was extinguished (AB-) in the second stage. At test, Rescorla observed less conditioned responding to a compound of AD than to BC. Rescorla concluded that, consistent with the prediction error principle, the associative strength of A declined more than that of B on the AB- trials. A different result is predicted by the theory protection account proposed in this thesis. This approach suggests that, using the same design, one should see the opposite result in humans to that observed in rats. This is because A+ training in the first stage shows that A must be a cause of the outcome. Participants would be expected to protect this theory from change. In contrast, B-

training in the first stage allows participants to ascertain with confidence that B is not a cause of the outcome, but does not allow them to confidently determine whether B is neutral or preventative of the outcome. Participants can protect their existing theories about A and B during AB- trials by concluding that B is preventative of the outcome that would otherwise have been produced by A. On the compound test, the theory protection account therefore predicts that the BC compound should receive lower causal ratings than AD.

The current chapter contains four experiments of this type, in which human participants were trained with a causal scenario that was again designed to allow the comparison of the prediction error and theory protection accounts. Experiments 6 and 7 also tested the theory protection account against both the Mackintosh (1975) and the Pearce and Hall (1980) attentional accounts. In the second training stage of each experiment, a previously trained causal cue was trained in compound with another cue with a more ambiguous causal status, including a close replication of the Rescorla (2001) design in Experiment 3. It is worth noting that a variant of this design was conducted by Haselgrove and Evans (2010), although these were some important methodological differences. Their design and findings are discussed in more detail in the General Discussion (3.6) at the end of this chapter. A key difference, in comparison to the Chapter 2 experiments, was that the compounds were followed by the absence of the outcome (rather than the presence of the outcome). This means that these experiments used an extinction design rather than an acquisition design. Furthermore, the types of cue used to create differing degrees of confidence were very different to those used in the Chapter 2 experiments. The scenario employed was also different. Based on the

theory protection account, it was expected that there would be more learning about the ambiguous cue during compound trials.

The experiments reported in this chapter also encompass some other important differences to those in Chapter 2. For example, the suggestion about cues given high and low causal ratings differing in their rate of associability chance (see Chapter 2.5: General Discussion) is tested. Participants were trained with one cue that was a cause of an outcome, which also had a non-ambiguous causal status, alongside another cue that did not cause that outcome, but which was less predictive and had a more ambiguous status. In other words, the causally certain cue would have higher ratings (after the initial training stage) than the causally uncertain cue. This is the opposite to the experiments in the previous chapter. Additionally, Experiment 7 further tested the theory protection account by asking participants to give their confidence about the causal status of each cue (in the form of a Probe Test and a Forced Choice stage), prior to the compound training stage. It is also worth noting that Experiments 6 and 7 predicted more learning about a cue with no apparent prediction error, than one with a large prediction error. Such a result is counter-intuitive, if one follows the basic logic of learning being proportional to prediction error. In other words, if there is no prediction error, then there should be no learning.

3.2: Experiment 4

This experiment used an allergist task in which participants had to learn whether an allergic reaction would occur, on the basis of different chemicals being ingested. Participants were presented with a fictional scenario in which they were working in a drug research setting, trying to work out which chemicals cause the side effect of a stomach ache in a test patient. Participants were told that chemicals could be causal, neutral, or preventative with respect to the stomach ache outcome. On each training trial, participants were presented with one or more chemicals and were asked to predict whether or not the patient would experience a stomach ache after consuming them. Once participants had made their prediction, they were provided with feedback as to whether or not a stomach ache occurred. Following training, participants were tested by being asked to make ratings indicating how likely they thought a stomach ache would be after the patient ingested specific chemicals singly or in pairs. The design of the experiment is shown in Table 2. Participants were initially trained with two causal cues (A+ and C+), and two non-causal cues (E- and F-). The non-causal cues were added as fillers, so that participants experienced both the presence and the absence of the stomach ache during Stage 1. Next, participants were trained with a non-causal compound AB- that consisted of a previously causal cue (A) and a novel cue (B). The novel cue was chosen to minimize the extent to which participants might have any causal theory about B at the start of AB- training. According to individual prediction error there should be more learning about A, since it was consistently paired with the outcome during Stage 1. According to overall prediction error there should be equal learning about A and B, because both of these cues share their prediction error. However, the theory protection account predicts more learning about B, since participants should protect their theory that A is causal, and instead attribute the absence of the outcome to the novel B.

Table 2. The design of experiments 4-6

Experiment	Stage 1	Stage 2	Test
4	A+ C+ E- F-	AB-	AD BC A B C D E F
5	A+ C+ E- F- B? D?	AB-	AD BC A B C D E F
6	A+ C+ B- D-	AB-	AD BC A C B D

Key:
Letters A-F = different cues
+ = stomach ache
- = no stomach ache
? = outcome concealed from participants

Learning about A and B was compared using a final test discrimination equivalent to Rescorla (2001). In addition to being asked for the likelihood of the outcome for each individual cue, participants were asked to give likelihood ratings for two compounds, AD and BC. Each of these compounds contained one cue that had been trained as causal in Stage 1, and one that had not been encountered in Stage 1. In the absence of Stage 2 training, these two compounds should have been assigned the same likelihood ratings at test. Consequently, any difference between these compounds must necessarily have been the result of the AB- training in Stage 2, and would indicate differing amounts of learning about A and B during that stage. If learning is governed by individual prediction error, participants should have rated the likelihood of the outcome as being lower for AD than BC, as a result of A decreasing in associative strength during the AB- trials. If learning is governed by overall prediction error, then there should have been no difference between ratings of AD and BC. However, if learning is determined by the theory protection account, then BC should have been assigned lower ratings than AD at

test, indicating that participants learned during the AB- trials that B is preventative of the outcome, in order to protect their existing theory that A is causal.

3.2.2: Method

Participants

Forty psychology students from the University of Plymouth participated in this experiment, in return for course credits (28 female, 11 male, 1 non-binary; mean age = 23.05 , SD = 7.03). This sample size has adequate power to detect medium-sized within-subjects effects (87% power at d = 0.5). People who had previously taken part in similar experiments were excluded from this study, to ensure participants were naive to the purpose of the experiment.

Materials

Participants were tested in the same lab at the University of Plymouth. The experiment was conducted using Viglen Genie desktop computers, running the Windows 10 operating system. The computers all used 22-inch Phillips LED displays, with participants at a typical distance (of approximately 40-80 cm) from the screen. The experiment was designed and executed in Psychopy (Peirce, 2007), with the output generated as individual CSV files for each participant. Participants made their responses by pressing keys on a standard UK computer keyboard during the training stages, and by using mouse clicks during the test stage. The six individual cues were represented on screen as different coloured images of shapes: blue oval, green square, grey triangle, pink diamond, purple star, yellow circle. All the coloured shapes were presented within a white square. The dimensions of each cue (including the white square) were 300 x 300 pixels on a 1920 x 1080 pixel screen. For each participant, the coloured shapes were randomly assigned to serve as A, B, C, D, E, and F. The two outcomes, ‘stomach ache’

and ‘no stomach ache’, were represented by text on screen and a photograph of a man clutching his stomach, or a man giving a ‘thumbs up’, respectively. The outcome images were presented within a white rectangle. The dimensions of the outcome images (including the white rectangle) were 291 x 332 pixels. All experimental text, including instructions, was white. A black background was used throughout the experiment. Study information sheets, consent forms and debrief forms were all printed on paper.

Design

The experiment used a within-subjects design, as outlined in Table 2. During Stage 1, participants were presented with twelve blocks of training. The four trial types (A+, C+, E-, F-) appeared in a random order within each block. Each trial type was only presented once within each block. There were six trials in Stage 2, all with AB-. During the Test stage, participants were presented with two blocks of test cues. The eight trial types (A, B, C, D, E, F, AD, BC) appeared in a random order within each block. Each trial type was only presented once within each block.

Procedure

Participants were required to read an information sheet and sign a consent form prior to participating in the experiment. The experimental instructions were presented on the screen at the start of the experiment. They were adapted from Spicer et al. (2019) and displayed as follows:

This study is concerned with the way in which people learn about relationships between events. In the present case, you should learn whether the consumption of chemicals used in drug research, leads to an allergic reaction.

Imagine that you are working in a drug research laboratory, studying chemicals for potential use in medication. You are trying to identify which chemicals cause the side effect of a stomach ache in your test patient.

To identify which chemicals they react to, the patient ingests specific chemicals and observes whether a stomach ache occurs or not. The results of these tests are shown to you on the screen one after the other.

You will then be asked to predict whether the patient suffers from stomach ache. For this prediction, please click on the appropriate response button. After you have made your prediction, you will be informed whether your patient suffered from stomach ache or not. Use this feedback to find out which chemicals cause a stomach ache in your patient.

You will always be told what your patient has ingested. Sometimes, they have only consumed a single chemical and other times they have consumed two different chemicals. Please look at the chemicals carefully.

At first you will have to guess the outcome because you do not know anything about your patient. But eventually you will learn which chemicals lead to stomach ache in this patient and you will be able to make correct predictions.

All of the chemicals being studied can be easily identified by a unique logo. Each logo is both a different shape and a different colour.

For all of your answers, accuracy rather than speed is essential. Please do not take any notes during the experiment. If you have any questions, please ask them now.

Please note, some chemicals will cause a stomach ache, while others will be neutral and will not cause a stomach ache. However, it is also possible for specific chemicals to actively PREVENT a stomach ache from occurring in your patient.

If you do not have any questions, please start the experiment by pressing the space bar.

For each trial during the training stages, the cues were visually presented on either the left-hand side of the screen or the right-hand side of the screen. When only one image was presented, the opposite side of the screen contained a blank space. The cues were randomly assigned to either the left or right position on each trial. Text at the top of the screen stated that ‘The patient ingests the following:’, with the cues presented below this. Underneath the cues, further text stated ‘Which outcome do you expect? Please use your keyboard to respond’. Participants were instructed to respond by pressing the appropriate key on their keyboard; Z for ‘No Stomach Ache’ and M for ‘Stomach Ache’. After participants made their response, the feedback for that trial was shown. The feedback screen consisted of the appropriate outcome image along with its accompanying text, indicating either ‘Stomach Ache’ or ‘No Stomach Ache’. The feedback was shown on screen for two seconds, after which the next trial began.

After the completion of Stage 1, Stage 2 started with no trial break, so that from the perspective of participants this was a seamless continuation of the training. Stage 2 consisted of a previously unseen compound AB- presented six times in a row. As in Stage 1, the cues were randomised on each trial to appear on either the left- or right-hand side of the screen. The on-screen text and responding via the keyboard was the same as in Stage 1. The process for displaying the trial feedback was also the same, except that all six trials resulted in 'Stomach Ache' as the outcome.

After the completion of Stage 2, a further instruction screen was shown before commencement of the Test stage:

Next, your task is to judge the probability with which specific chemicals cause stomach ache in your patient. Single chemicals and pairs of chemicals will be shown to you on the screen.

In this part of the experiment, you will receive no feedback about the actual reaction of the patient. Use the information that you have collected so far, to make your rating.

Press space bar to continue the experiment.

For each trial during the Test stage, the cues were visually presented on either the left- or right-hand side of the screen. When only one image was presented, the opposite side of the screen again contained a blank space. The cues were randomly assigned to either the left or right position on each trial. As before, text at the top of the screen stated that

‘The patient ingests the following:’, with the cues presented below this. Underneath the cues, further text stated ‘How likely are they to suffer a stomach ache? (0 = Very Unlikely; 10 = Very Likely)’. Participants were instructed to respond by clicking on an 11-point rating scale using their mouse pointer, to indicate how likely they thought the occurrence of a stomach ache would be. The rating scale was located in the lower part of the screen, with the 11-point scale running from left to right, in ascending numerical order. After participants made their response, a blank screen appeared for 0.4 secs, after which the next test trial was presented. Following the completion of the experiment, participants were provided with a debrief form.

Analysis

The data were processed and analysed using R (R Core Team, 2018). The difference between the AD and BC test compounds was assessed using paired-samples t-tests. Some additional analyses were also conducted on key single test stimuli, to test for specific predicted differences between cue ratings. The alpha level was set to $p < .05$ for all tests. As these tests were done on the basis of specific prior predictions, there was no need to correct for multiple comparisons. Bayesian t-tests were also conducted, using the procedure recommended by Dienes (2011) and implemented as R code by Baguley and Kaye (2010). A uniform distribution was specified as the prior for each test, with a lower limit of -10 and an upper limit of 10 (in terms of the mean difference between ratings), because these are the largest mean differences in either direction permitted by an 11-point test rating scale. In keeping with accepted conventions (e.g. Jeffreys, 1961), a Bayes factor of over three was set as the level providing evidence for a difference,

while a Bayes factor of less than one third was set as the level providing evidence for no difference. Values between these levels were accepted as being inconclusive.

3.2.3: Results and Discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/bzerh/. The descriptive statistics for the Experiment 4 training stages are shown in Figure 7. The data from Stage 1 indicate that participants learned sufficiently about the four different cues by the time Stage 1 was complete. Similarly, the data from Stage 2 indicate that participants learned that the AB- compound was non-causal by the end of Stage 2.

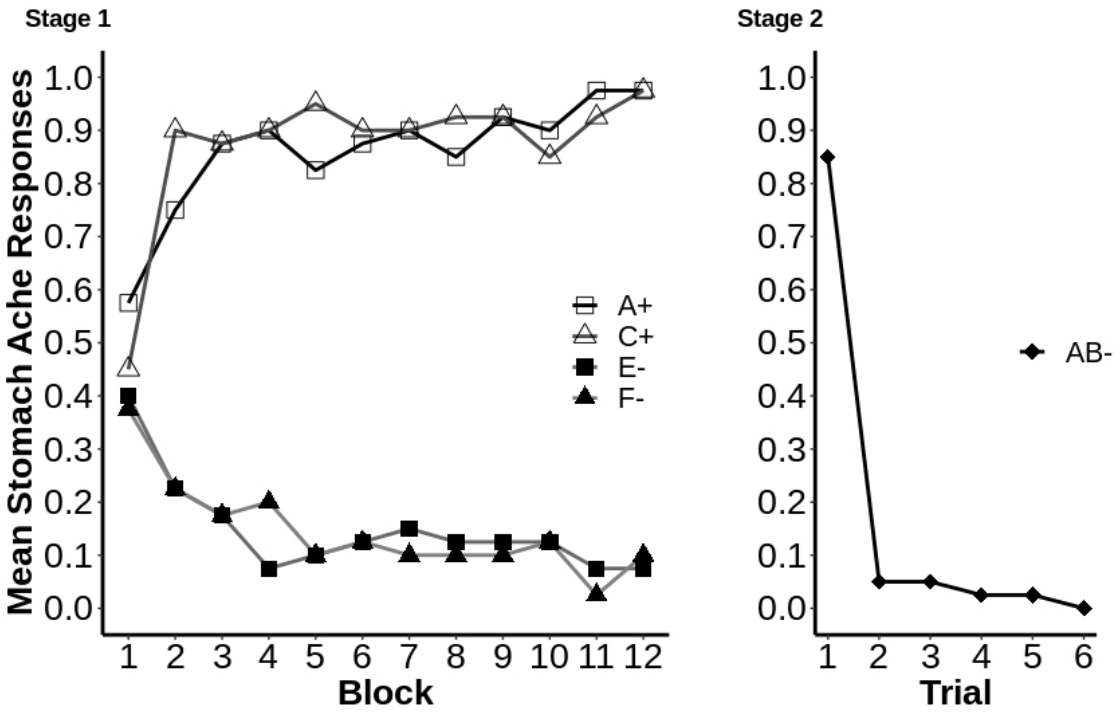

Figure 7. Experiment 4 training Stage 1 and 2 data.

Descriptive statistics for the Experiment 4 Test stage are shown in Figure 8. Ratings for BC were significantly lower than for AD; $t(39) = 5.61$, $p < .001$, $BF = 4.08 \times 10^5$, $d = .89$. Further testing revealed lower ratings for the single cue B compared to D; $t(39) = 8.48$, $p < .001$, $BF = 2.23 \times 10^{14}$, $d = 1.34$. Conversely, there was no difference between the ratings assigned to A and C; $t(39) = .04$, $p = .919$, $BF = .05$, $d = .02$, indicating that

participants did not learn about A during Stage 2. Taken as a whole, these findings show that the differences between the compounds were specifically driven by learning about B during AB- trials, with participants maintaining their association between A and the occurrence of stomach ache. Figure 3 panel B shows inter-subject variability on the key compound test difference. As would be expected from the mean test ratings, most individual participants rated BC lower than AB, although the variability did extend to some participants assigning a lower rating to AB. It is also worth noting that the intermediate mean rating for D during the Test stage is consistent with the idea that participants do not know the causal status of novel cues, and are consequently unlikely to assign high or low ratings (see Chapter 4 four a detailed exploration of uncertainty about novel cues). These data appear to support the theory protection account, as opposed to a prediction error account (Rescorla, 2001). However, one possible limitation of this experiment is that novel cues are often regarded as being more salient than familiar cues (e.g. Lubow & Moore, 1959). Consequently, B may have been more salient than A during training Stage 2, and this may have led to more learning about B than A, despite B having a smaller prediction error. Experiment 5 addressed this alternative account, by replacing the two novel cues with familiar cues that had a causally unknown status.

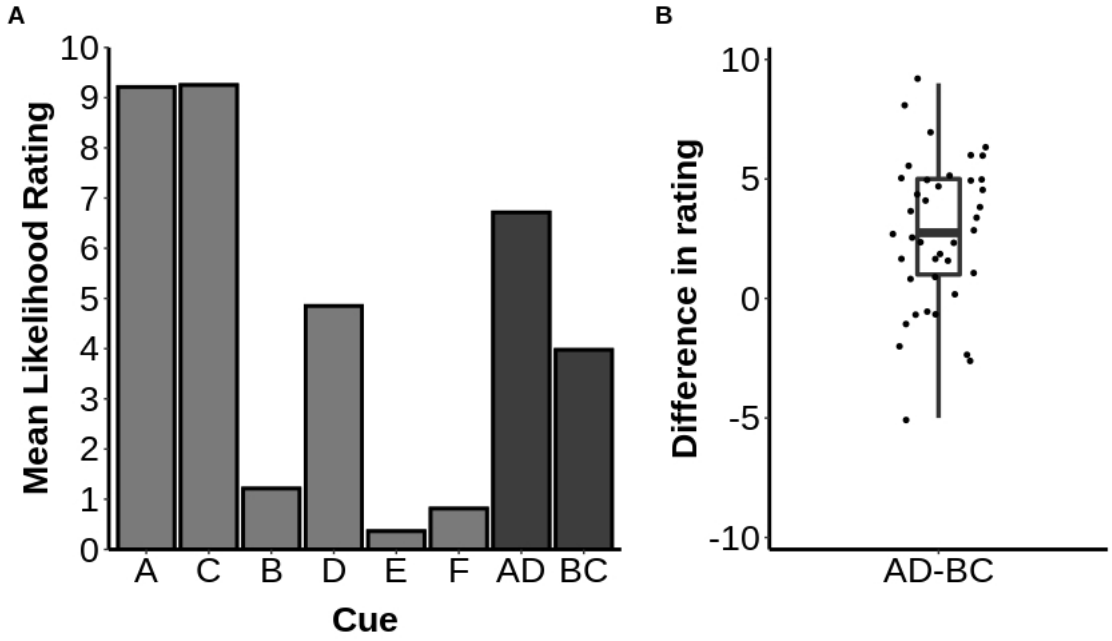

Figure 8. Panel A shows Experiment 4 Test stage ratings for single cues and compounds. Panel B is a plot showing inter-subject variability on the key AD-BC difference.

3.3: Experiment 5

This was a variant of Experiment 4 in which cues B and D were causally ambiguous but not novel. This was achieved by exposing participants to cues B and D during Stage 1, but not revealing their causal status. In keeping with the allergy paradigm, participants were informed that the patient information was missing on these trials, rather than being told whether or not a stomach ache occurred. This allowed B and D to be familiar at the end of Stage 1 of the experiment, but for participants to still hold no strong theory about their causal status. It also permitted an examination of the generality of Experiment 4 by testing whether the theory protection account would still apply when a different type of causally-ambiguous cue was employed. The design of Experiment 5 is shown in Table 2. All of the other experimental details were the same as Experiment 4. The experimental predictions were also the same, in that cue A would have the greater prediction error at the start of Stage 2, and participants would be less likely to have a theory about B. As before, the expectation was that participants would maintain their belief that A is a cause during Stage 2, instead learning about B, and giving lower ratings for the BC compound at test.

3.3.2: Method

Participants

Thirty-six psychology students from the University of Plymouth participated in this experiment, in return for course credit (26 female, 10 male; mean age = 24.67, SD = 6.81). This sample size has adequate power to detect medium-sized within-subjects effects (83% power at d = 0.5). People who had previously taken part in similar experiments were excluded from this study, to ensure participants were naive to the purpose of the experiment.

Materials

The materials used for Experiment 5 were the same as those used for Experiment 4, with the exception of an additional image and text used for the concealed-outcome trials. The ‘information missing’ trial feedback was represented by text on screen and an image of a black question mark. As with the other trial feedback, the image was presented within a white rectangle. The dimensions of the image (including the white rectangle) were 291 x 332 pixels.

Design

The experiment used a within-subjects design, as outlined in Table 2. During Stage 1, participants were presented with twelve blocks of training. The six trial types (A+, B?, C+, D?, E-, F-) appeared in a random order within each block. Each trial type was only

presented once within each block. Stage 2 and the Test stage were identical to Experiment 4.

Procedure and analysis

Apart from the changes described above, the procedure and analysis for Experiment 5 were the same as for Experiment 4.

3.3.3: Results and Discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/amubk/. The descriptive statistics for the Experiment 5 training stages are shown in Figure 9. The data from Stage 1 indicate that participants learned sufficiently about the six different trial types by the time training was complete. The intermediate responses for B and D are consistent with participants being unconfident about the causal status of these cues. Similarly, the data from Stage 2 indicate that participants learned that the AB- compound was non-causal by the time training was complete.

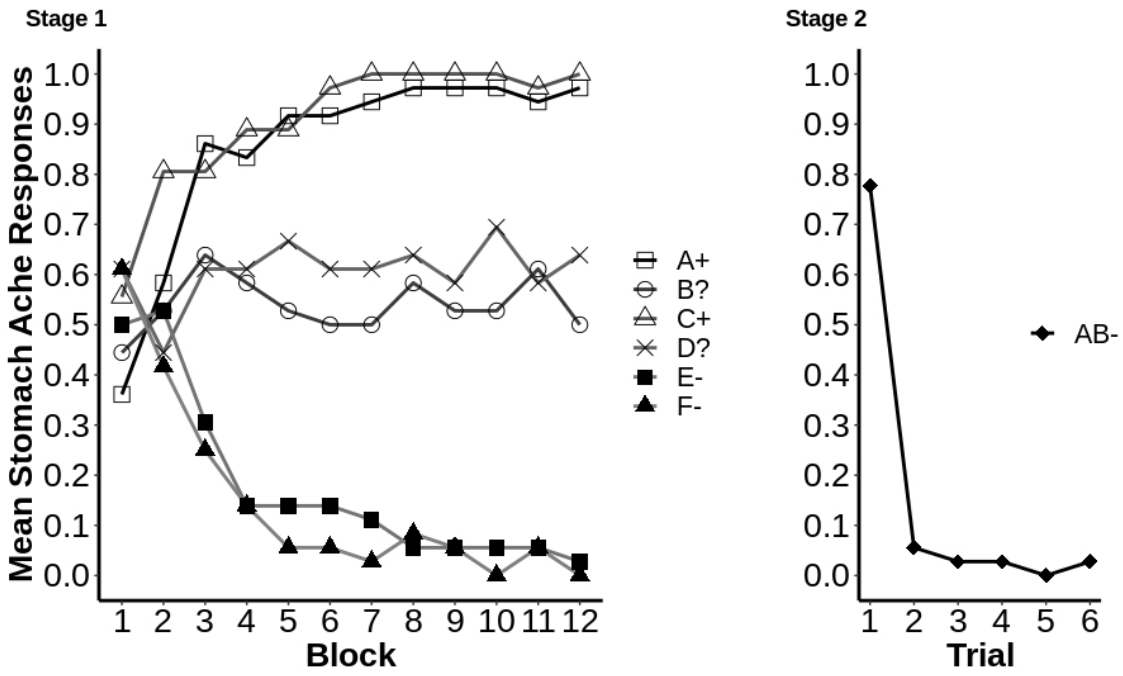

Figure 9. Experiment 5 training Stage 1 and Stage 2 data.

Descriptive statistics for the Experiment 5 Test stage are shown in Figure 10. Ratings for BC were significantly lower than for AD; $t(35) = 3.34$, $p = .002$, $BF = 17.83$, $d = .56$. Further testing revealed lower ratings for the single cue B compared to D; $t(35) = 5.06$, $p < .001$, $BF = 2.17 \times 10^4$, $d = .84$. Conversely, there was no difference between the

ratings assigned to A and C; $t(35) = .50$, $p = .62$, $BF = .03$, $d = .08$. These findings indicate that the differences between the compounds were specifically driven by learning about B during AB- trials. As with the novel cues in Experiment 4, the intermediate rating for D at test is consistent with participants holding no reliable theory about the causal status of concealed-outcome cues. Taken as a whole, these data are again consistent with the theory protection account, as opposed to a prediction error account (Rescorla, 2001). This is because the novel and concealed-outcome cues had a smaller prediction error than the causal cues at the start of Stage 2 in both experiments, as indicated by the intermediate responses for B and D during Stage 1 (for the concealed-outcome cues) and the intermediate ratings for D during the Test stage (for both the novel cues and the concealed-outcome cues).

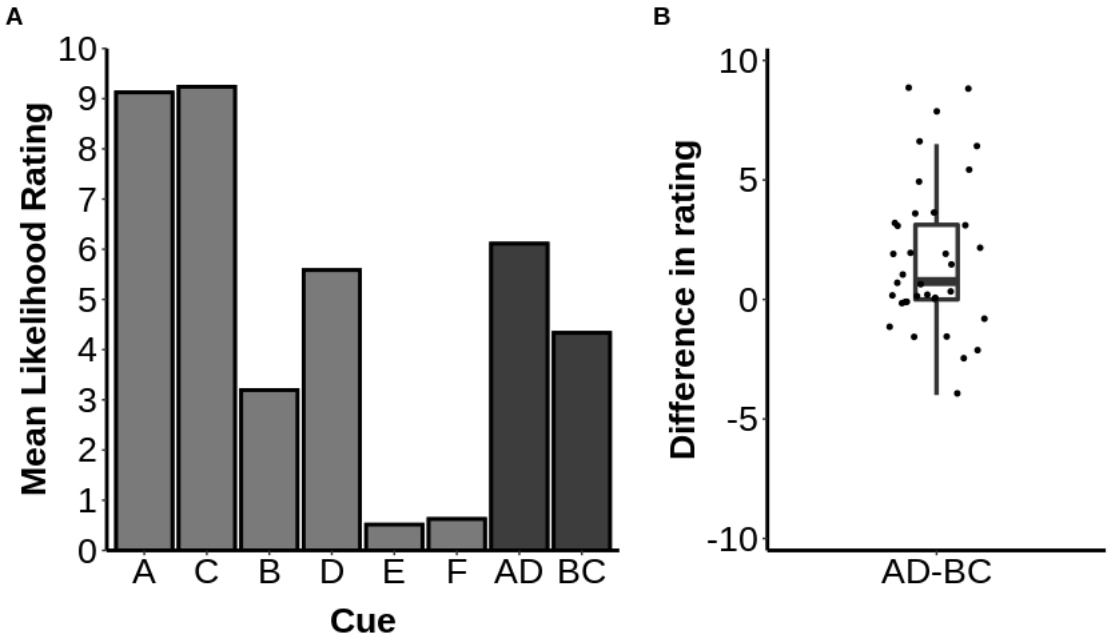

Figure 10. Panel A shows Experiment 5 Test stage ratings for single cues and compounds. Panel B is a plot showing inter-subject variability on the key AD-BC difference.

Whilst the Experiment 4 and 5 data are irreconcilable with a prediction error account, they could be explained according to the attentional processes represented Pearce and Hall (1980) and Mackintosh (1975) models. Recall that, according to Pearce and Hall (1980), associability is highest for cues that are unreliable predictors of outcomes, which results in greater learning about those cues. Importantly, novel cues are stated to have high starting associability, which then declines if these cues are reliable predictors. In Experiment 4, the associability of A should have declined during Stage 1, because it was a good predictor of stomach ache. Meanwhile, the associability of B should have been high at the start of Stage 2, as a consequence of its novelty. Pearce and Hall's model therefore predicts more learning about B than A during the AB- trials, at least at the start of Stage 2. In Experiment 5, the associability of B should have remained high across Stage 1, despite its familiarity, because participants did not have an opportunity to observe the outcome. A, meanwhile, should have declined in associability as a consequence of being consistently paired with stomach ache. Hence, Pearce and Hall's model provides a plausible account of both Experiments 4 and 5. Of course, it has already been established that the Pearce and Hall attentional account is not able to capture the results of the experiments reported in Chapter 2. However, a further investigation of this account, in the context of the simpler experimental designs in this Chapter seemed worthwhile.

Recall also that according to Mackintosh's (1975) model, the best available predictor of each trial's outcome increases in associability, while other cues decrease, with subsequent learning being the product of this updated associability and an individual prediction error. Similarly to Experiment 3, the way in which Mackintosh's (1975) model can capture the Experiment 4 and 5 results is more subtle than Pearce and Hall's

(1980) explanation outlined above. This is because, during Stage 1 of Experiment 4, A should have received an increase in associability. As a result, the associability of A should have been higher than that of B at the start of Stage 2. The same is perhaps true of Experiment 5, since withholding information about the consequences of B might have prevented participants from learning that B was predictive. A should therefore have had both the greater individual prediction error and the higher associability at the end of Stage 1 in both experiments. This means that A should have been learned about the most at the start of Stage 2. However, after the first Stage 2 trial, the outcome predicted by B would have been less discrepant with the AB- outcome (i.e. no stomach ache) than the outcome predicted by A. Hence, the associability of B would have increased, while the associability of A would have decreased. These changes in associability could produce more learning about B, in spite of it having the smaller prediction error throughout Stage 2. Since both Mackintosh's and Pearce and Hall's (1980) model can be reconciled with the results of Experiments 4 and 5, Experiment 6 was designed such that both of these models make different predictions to the theory protection account.

3.4: Experiment 6

This experiment was a close replication of the Rescorla (2001) rat design, also using a very similar design to that of Experiments 4 and 5 to test the theory protection account against Pearce and Hall's (1980) and Mackintosh's (1975) models. The design of Experiment 6 is shown in Table 2. To ensure that Pearce and Hall's model made a different prediction to the theory protection account, cue B was a good predictor of the trial outcome during Stage 1. In place of a novel cue (Experiment 4) or a cue with a concealed outcome (Experiment 5), here B was reliably followed by the absence of stomach ache. According to Pearce and Hall, since A and B were equally predictive during Stage 1 (A of stomach ache and B of no stomach ache), they should both have suffered equivalent decrements in associability during Stage 1 and there should have been no difference in learning about these cues during Stage 2. This design also ensured that there was no individual prediction error associated with B on the Stage 2 AB- trials, since both these trials and the preceding Stage 1 trials with B ended in the absence of stomach ache. Mackintosh's model therefore predicts no learning about B in Stage 2, because it proposes that learning is proportional to individual prediction error. As with the previous experiments, Rescorla's (2001) account predicts more learning about A than B during Stage 2 because of its larger individual prediction error.

The theory protection account predicts more learning about B than A on AB- trials, just as in Experiments 4 and 5. This is because, following B- training, the status of B should still have been somewhat ambiguous; it might have been neutral with regard to stomach ache, or it might have been preventative. Therefore, to protect their theory about cue A being causal, participants were expected to infer that cue B was preventative. The use of

B- and D- trials in Stage 1 had one other minor implication for the experimental design. In this experiment, the E- and F- trials were not needed to provide experience of non-reinforced cues in Stage 1, as in the previous experiments. As before, the Test stage contained two compounds, AD and BC, which permitted a comparison of learning about A and B. Based on the theory protection account, it was expected that there would be higher ratings for AD than for BC. This is in contrast to the predictions of other models, which either predict no difference (Pearce and Hall, 1980), or higher ratings for BC (Mackintosh, 1975; Rescorla, 2001). Because Stage 1 training should have at least allowed participants to learn with confidence that B was not a cause of stomach ache (unlike in Experiments 4 and 5), it was expected that the size of the AD-BC difference would be smaller than in previous experiments.

3.4.2: Method

Participants

Forty psychology students from the University of Plymouth participated in this experiment, in return for course credit (35 female, 5 male; mean age = 22.03 SD = 7.55). This sample size has adequate power to detect effects somewhat smaller than those observed in Experiments 4-5. Specifically, the mean effect size for the AD-BC comparison across these two experiments is d = 0.73. At the current sample size, a 38% reduction in that effect size (to d= 0.45) would still result in adequate (80%) power. People who had previously taken part in similar experiments were excluded, to ensure participants were naive to the purpose of the experiment.

Materials

The materials used for Experiment 6 were the same as those used for Experiment 4.

Design

The experiment used a within-subjects design, as outlined in Table 1. During Stage 1, participants were presented with twelve blocks of training. The four trial types (A+, B-, C+, D-) appeared in a random order within each block. Each trial type was only presented once within each block. There were six trials in Stage 2, all with AB-. During the Test stage, participants were presented with two blocks of test cues. The six trial types (A, B, C, D, AD, BC) appeared in a random order within each block. Each trial type was only presented once within each block.

Procedure and analysis

Apart from the changes described above, the procedure for Experiment 6 was the same as for Experiments 4 and 5. The analyses were the same, except that the Bayesian priors for the compound test were updated on the basis of the mean result across Experiments 4 and 5. As explained above, in Stage 1 of Experiment 6, the range of outcomes that could be caused by cue B was reduced, so there was a theoretical basis for expecting the mean test difference to be smaller than the values observed in Experiments 4-5. Therefore, following the recommendations of Dienes (2011), a half-normal prior distribution was specified for the Bayesian t-test on the compounds, with a mean of zero and a standard deviation set to the corresponding mean difference observed across Experiments 4-5. A uniform prior, equivalent to Experiments 4-5, was specified for the Bayesian t-tests on single cues. This is because neutral and preventative cues should both be given a low rating on the 11-point likelihood scale, so no difference was anticipated for either of the tests on single cues.

3.4.3: Results and Discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/vqnbc/. Descriptive statistics for the training stages are shown in Figure 11. The data from Stage 1 indicate that participants learned sufficiently about the four different trial types by the time training was complete. Similarly, the data from Stage 2 indicate that participants learned that the AB- compound was non-causal by the end of Stage 2. The response data for the first Stage 2 trial is consistent with the participants lacking confidence as to whether B was neutral or inhibitory.

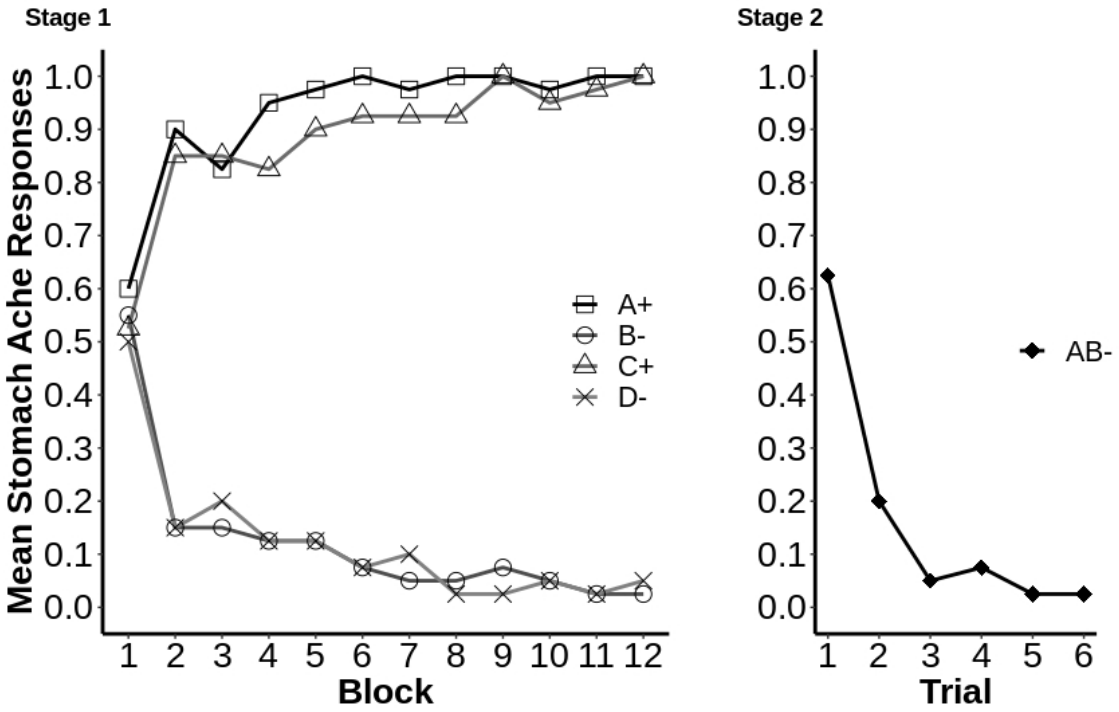

Figure 11. Experiment 6 training Stage 1 and 2 data.

Descriptive statistics for the Test stage are shown in Figure 12. Ratings for BC were significantly lower than for AD; $t(39) = 3.37$, $p = .002$, $BF = 89.84$, $d = .53$. The eleven-point rating scale did not allow any difference between neutral and preventative cues to

be detected, since both would be given a low rating. Therefore, there was no reason to expect ratings for B and D to differ $t(39) = .54$, $p < .59$, $BF = .05$, $d = .09$. As with Experiments 1-2, there was no difference between the ratings assigned to A and C; $t(39) = .50$, $p = .62$, $BF = .02$, $d = .21$. These data are again consistent with the theory protection account and are, like the results of Experiments 4 and 5, the opposite of the outcome anticipated by an individual prediction error account. Notably, the results confirm the prediction that there would be greater learning about a cue with no apparent prediction error, than one with a large prediction error. Crucially, neither Pearce and Hall's (1980) nor Mackintosh's (1975) attentional models are able to account for these findings either. According to Pearce and Hall, learning about B during Stage 2 should be no greater than that for A, since both cues should have an equivalent decline in associability during Stage 1. The lack of any individual prediction error with respect to cue B, on the AB- trials of Stage 2, also means that Mackintosh's model predicts no learning for this cue. Cue A, by contrast, whatever its associability, would be expected to extinguish on AB- trials. Hence, Mackintosh predicts that compound AD will be assigned lower ratings than BC on test, which is the opposite of the observed result.

According to the theory protection account, participants should have lacked confidence about the causal status of B at the end of Stage 1, in spite of learning that it was not a cause. However, the results of Experiment 6 contain no explicit evidence that this is so. This is because the response method used to predict the presence or absence of the outcome did not allow participants to express any lack of confidence about whether a cue was neutral or preventative. Although the responses made on the first Stage 2 trial might indicate some lack of confidence about whether a stomach ache would occur (see Figure 11), this could be due to either lacking causal confidence about B or some more

general causal uncertainty about the novel combination of cues. Experiment 7 was intended to address this issue by including a measurement of participants' relative confidence about A and B at the end of Stage 1. Experiment 7 also tested whether participants had any sense that cue B might be to some extent inhibitory, as predicted.

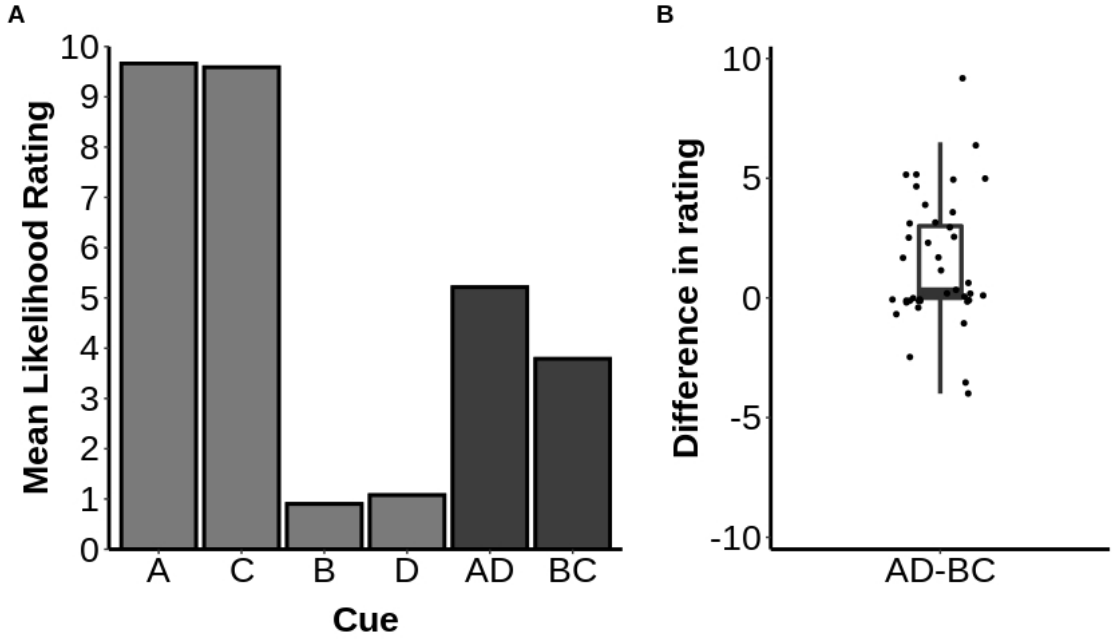

Figure 12. Panel A shows Experiment 6 Test stage ratings for single cues and compounds. Panel B is a plot showing inter-subject variability on the key AD-BC difference.

3.5: Experiment 7

In addition to replicating the results of Experiment 6, the aim of Experiment 7 was to assess whether there was a difference in participants' confidence about the causal status of A and B before the start of Stage 2. The design was identical to that of Experiment 6, except for the addition of two extra stages between Stage 1 and Stage 2. The full design of Experiment 7 is shown in Table 3. The first addition was a Probe Test, in which participants were asked to make ratings about cues A and B. Unlike the Test stage ratings, for which participants were asked to rate the likelihood of the outcome given a specific cue or compound, participants were instead asked what they thought each of cues A and B did. Participants were presented with a 21-point scale, running from -10 to +10, where +10 meant the cue causes the outcome, 0 meant the cue is neutral, and -10 meant the cue prevents the outcome. The second addition was a Forced Choice stage, in which participants were asked to choose which of those two ratings (i.e. the ratings assigned to A and B on the 21-point scale) they were most confident about. If participants were confident about the causal status of A, but comparatively unconfident about whether B was neutral or inhibitory, then they should have given a high rating to A on the positive end of the Probe Test scale, and an intermediate rating to B in the negative half of that scale. Furthermore, when asked which of these ratings they were most confident about during the Forced Choice stage, they should have chosen A. These findings would support the theory protection account.

Table 3. The design of Experiment 7

Experiment	Stage 1	Probe Test	Forced Choice	Stage 2	Test
4	A+ C+ B- D-	A B	A or B	AB-	AD BC A C B D

3.5.2: Method

Participants

Sixty-one psychology students from the University of Plymouth participated in this experiment, in return for course credit (51 female, 10 male; mean age = 21.02, SD = 6.01). This sample size has excellent power to detect the key AD versus BC comparison at the effect size observed in Experiment 3 (99% power at d = 0.53), excellent power to detect a medium-sized effect on the Probe Test (97% power at d = 0.5), and adequate power to detect a medium-to-large effect on the forced-choice test (80% power at w = 0.36). People who had previously taken part in similar experiments were excluded, to ensure participants were naive to the purpose of the experiment.

Materials

The materials used for Experiment 7 were the same as those used for Experiment 4 and Experiment 6.

Design

The experiment used a within-subjects design, as outlined in Table 3. Stage 1, Stage 2 and the Test stage used a design identical to Experiment 6. The Probe Test stage and Forced Choice stage were added between Stages 1 and 2. During the Probe Test, participants were presented with a single block containing only cues A and B. The cues appeared in a random order, and were only presented once within the block. The Forced Choice followed on immediately from the Probe Test. Participants were presented with

a single trial, in which cues A and B were presented together on screen. The screen position of A and B was randomised and counterbalanced, so that an equal number of participants saw these cues in each of the two possible left-right configurations.

Procedure

The procedure for Experiment 7 was the same as for Experiment 6, except for the addition of the Probe Test and Forced Choice stages. Following the completion of Stage 1, an instruction screen was shown before the commencement of the Probe Test:

Next, your task is to make ratings about two of the chemicals you have learned about. Firstly, you will be asked what these chemicals do when ingested by your patient (e.g. cause stomach ache, prevent stomach ache, neutral). You will be able to make your response using a rating scale.

Once you have made your ratings, you will be asked which of those ratings you feel most confident (i.e. certain) about. You will be able to make a response using a key press.

In this part of the experiment, you will receive no feedback about the actual reaction of the patient. Use the information that you have collected so far, to make your choices.

Press space bar to continue the experiment.

For each trial during the Probe Test, the cues were visually presented in the centre of the screen. Text at the top of the screen stated ‘Consider the following chemical:’, with the cue presented below this. Underneath the cue, further text stated, ‘What does this chemical do when ingested by your patient? (-10 = Definitely Prevents Stomach Ache; 0 = Definitely Does Nothing; 10 = Definitely Causes Stomach Ache)’. Participants were instructed to respond by clicking on a 21-point rating scale using their mouse pointer, to indicate what they believed the causal status of the cue to be. The rating scale was located in the lower part of the screen, with the 21-point scale running from left to right, in ascending numerical order. After participants made their response, a blank screen appeared for 0.4 secs, after which the next Probe Test trial was presented. Following the completion of the Probe Test, the Forced Choice test commenced. During the single Forced Choice trial, the cues were visually presented on either the left- or right-hand side of the screen. The cues were randomly pre-assigned for each participant to the left and right positions. Text at the top of the screen stated ‘You gave these ratings to the two chemicals you were just asked about:’, with the cues presented below this and the Probe Test ratings presented directly beneath each of the respective cues. Underneath the ratings, further text stated ‘Which of these two ratings are you most confident (i.e. certain) about?’ Participants were instructed to use their computer keyboard to respond (Z for ‘Left Rating’ and M for ‘Right Rating’). After participants made their response, a blank screen appeared for 0.4 secs, after which an instruction screen for Stage 2 was presented.

Next, you will continue to learn about the chemicals used in drug research, as you did during the first part of the experiment. As before, the patient ingests specific chemicals

and observes whether a stomach ache occurs or not. The results of these tests are shown to you on the screen one after the other.

You will then be asked to predict whether the patient suffers from stomach ache. For this prediction, please click on the appropriate response button. After you have made your prediction, you will be informed whether your patient suffered from stomach ache or not.

Press the space bar to continue.

Following the presentation of these instructions, Stage 2 of the experiment commenced and the experiment continued, following the same procedure as Experiment 6.

Analysis

The analyses were the same as for Experiment 4, 5, and 6, except that the Bayesian priors were updated on the basis of the results of Experiment 6, with the exception of the Probe Test. Therefore, following the procedure recommended by Dienes (2011), a normal distribution was specified as the prior for the Bayesian t-test on the compounds, with the mean set to the corresponding Experiment 6 mean and the standard deviation set to half this value. As with Experiment 6, a uniform prior was specified for the tests on single cues, as no difference was anticipated. As there was no suitable previous study on which to specify a plausible predicted effect size for the Probe Test, a uniform distribution was specified, with a lower limit of -10 and an upper limit of 10 (for the mean difference of the ratings from zero, and the mean difference between the unsigned

ratings). These limits were chosen because 10 is the largest possible difference for these comparisons. The Forced Choice data were analysed with a chi-square test.

3.5.3: Results and Discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/kwzdr/. The descriptive statistics for the training stages are shown in Figure 13. The data from Stage 1 indicate that participants learned sufficiently about the four different trial types by the time training was complete. Similarly, the data from Stage 2 indicate that participants learned that the AB- compound was non-causal by the end of Stage 2. As before, the intermediate proportion of stomach ache responses on the first Stage 2 trial supports the idea that participants lack confidence about the causal status of B.

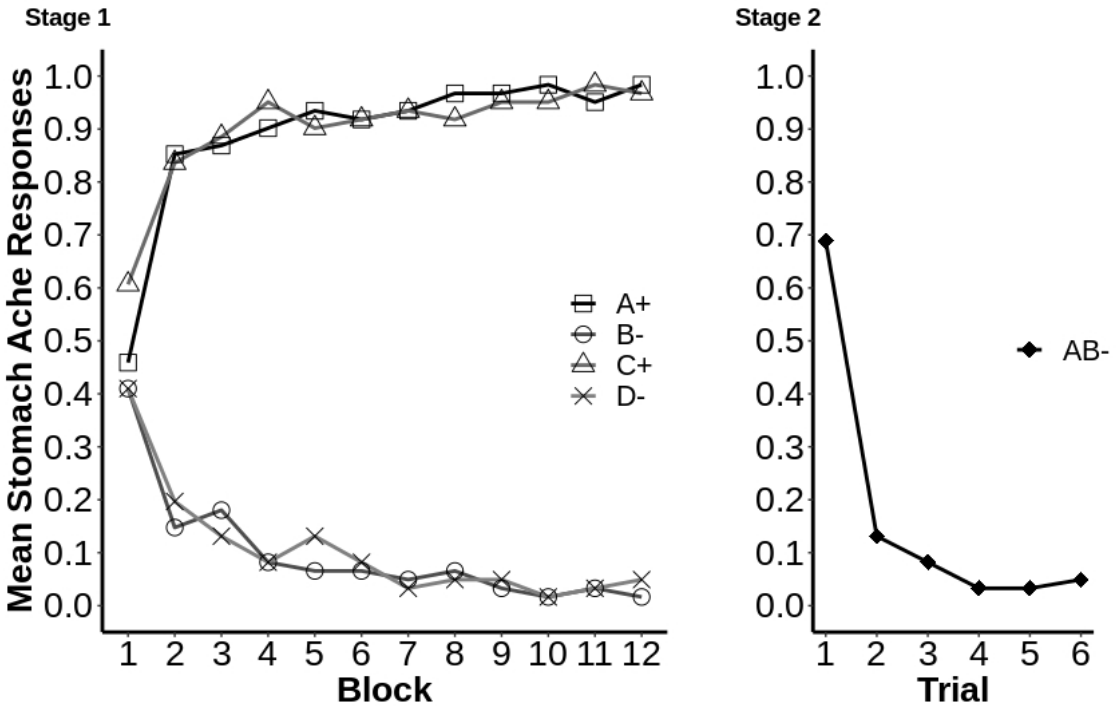

Figure 13. Experiment 7 training Stage 1 and 2 data.

Descriptive statistics for the final Test stage are shown in Figure 14. As in Experiment 6, ratings for BC were significantly lower than for AD; $t(60) = 4.89$, $p < .001$, $BF = 6.57$

x10^4, *d* = .63. As in Experiment 6, the 11-point Test stage rating scale did not allow any difference between neutral and preventative cues to be detected. Accordingly, there was no difference between the ratings assigned to B and D, *t*(60) = .97, *p* < .34, *BF* = 0.04, *d* = .12. There was also no difference between the ratings assigned to A and C; *t*(60) = 1.02, *p* = .31, *BF* = 0.04, *d* = .13, suggesting that nothing was learned about A during Stage 2. As a replication of Experiment 6, these findings provide further support to the theory protection account.

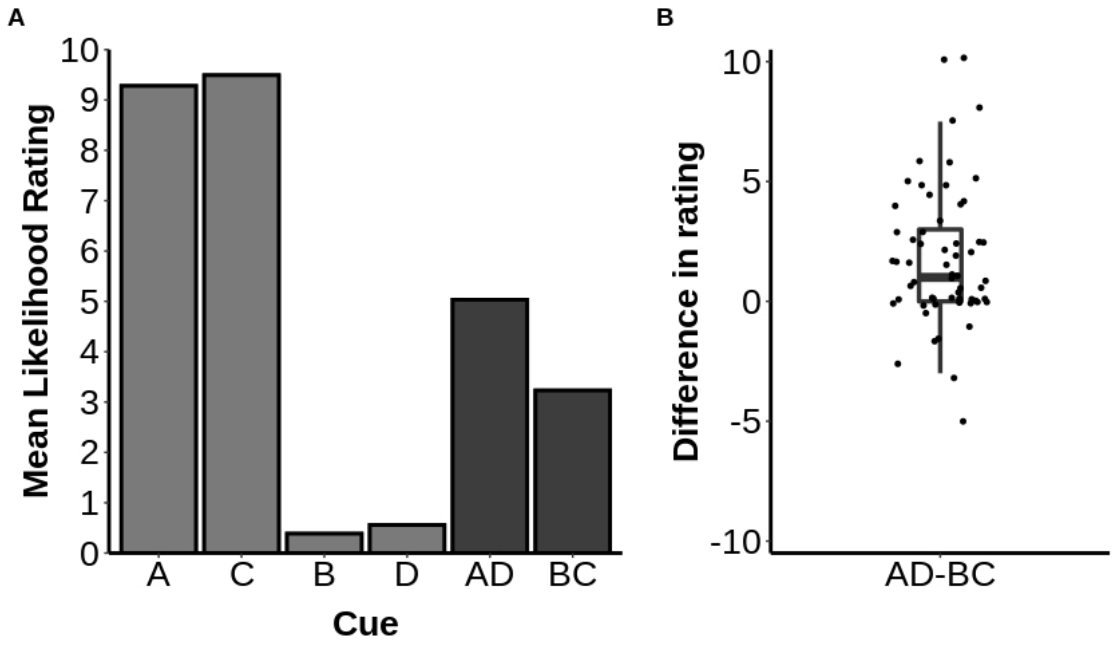

Figure 14. Panel A shows Experiment 7 Test stage ratings for single cues and compounds. Panel B is a plot showing inter-subject variability on the key AD-BC difference.

The descriptive statistics for the Experiment 4 Probe Test are shown in Figure 15. Ratings for A were significantly greater than zero; *t*(60) = 13.05, *p* < .001, *BF* = 7.55 x10^{35}, *d* = 1.67. Ratings for B were significantly below zero; *t*(60) = 7.89, *p* < .001, *BF* = 2.34 x10^{12} *d* = 1.01. The ratings for cue A were significantly further from zero than the ratings for cue B (i.e. the unsigned difference from zero was greater for A); *t*(60) = 8.56,

$p < .001$, $BF = 5.28 \times 10^{14}$, $d = 1.10$. The high positive mean rating for A, compared to intermediate negative mean rating for B, suggests greater confidence about the causal status of A than of B. These data support the prediction that participants would not know whether B was neutral or inhibitory. The Forced Choice results further support this view, since 51 participants chose their rating for cue A as the one they were most confident about, while only 10 chose B. A chi-square test demonstrated that this was significantly different to chance; $X^2(60) = 27.56$, $p < .001$, $w = .73$. These findings indicate that participants lacked confidence about the causal status of B prior to the start of Stage 2, despite having the opportunity to learn that it was not a cause. The results suggest that this lack of knowledge facilitated subsequent learning about B, while confidence about the causal status of A meant participants protected their theory about it, in spite of a larger prediction error. Deducing that B was preventative of stomach ache is consistent with B- trials in Stage 1, so participants could also protect their theory that B was not a cause.

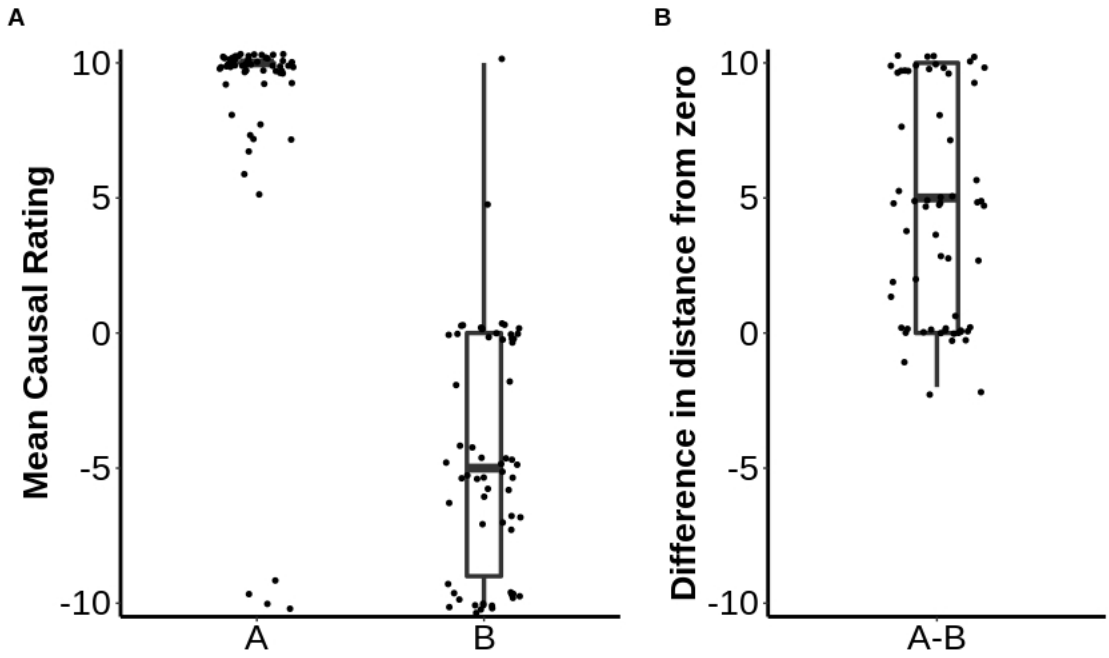

Figure 15. Panel A shows Experiment 7 Probe Test ratings for cues A and B. Panel B shows inter-subject variability on the difference between the unsigned differences from zero for these cues. Please note that the left hand boxplot in panel A is masked by the concentration of individual participants providing that cue with a high rating.

3.6: General Discussion

Taken as a whole, the findings of the four experiments in this chapter suggest that theory protection, in at least some circumstances, dictates the extent to which different cues are learned about when presented together. Participants appear to protect existing theories about the causal status of cues, instead attributing unexpected outcomes to cues about which they do not hold a strong theory. These findings cannot be accounted for by either individual (e.g. Bush & Mosteller, 1951) or overall (e.g. Rescorla & Wagner, 1972) prediction error, or a hybrid of both (e.g. Rescorla, 2001). This is because in each experiment there was greater learning about the cue with the smaller prediction error. Experiments 6 and 7 cannot be accounted for by Mackintosh's (1975) attentional model, since according to this model there should have been no prediction error for cue B during Stage 2, and hence no learning. Experiments 6 and 7 also cannot be accounted for by Pearce and Hall's (1980) attentional model, since the associability of both A and B should have been equally low at the end of Stage 1, resulting in equal learning during Stage 2. Furthermore, Experiment 7 provides direct evidence of a difference in participants' confidence before the compound training stage, supporting the idea that participants will protect their theory about a cue if they are confident about it. These data are consistent with the findings of Chapter 2, in their support of the theory protection account.

The experiments reported here also address the point raised in Chapter 2, in that cues given high and low causal ratings might differ in their associability changes, in the confines of the Mackintosh (1975) attentional account. In the three experiments reported in Chapter 2, a cue that was given low causal ratings during Stage 1, and a causally-ambiguous cue that was given intermediate causal ratings during Stage 1, were

presented in a compound during Stage 2. The compound was always followed by the outcome. In each experiment, participants learned more about the causally-ambiguous cue during Stage 2, apparently protecting their theory about the cues they had previously assigned low causal ratings. It was suggested that this could be a consequence of cues that are assigned lower ratings acquiring associability at a slower rate than those that are assigned higher ratings. This explanation is not viable for the experiments in Chapter 3, since the causally-ambiguous cue was always trained during Stage 2 with a cue that was predictive of stomach ache in Stage 1. Participants still learned more about the causally-ambiguous cue in each experiment, even though this cue was never a good predictor of stomach ache.

In the Chapter 2 General Discussion (2.5), differences were outlined between Pearce and Hall's (1980) model and the theory protection account. In the case of the theory protection account, future learning is influenced by knowledge about the cues, whilst in the case of Pearce and Hall's model, learning is influenced by surprisingness of the outcome (as this governs any update in associability for all cues that are present). This outcome-directed nature of Pearce and Hall's model is what prevents it from explaining Experiments 6 and 7, since this makes it insensitive to the negative associative strength acquired by B during Stage 1, which is central to the difference in learning (about A and B) according to the theory protection account. If one interprets the negative intermediate Probe Test rating in Experiment 7 as a lack of confidence about B being neutral or preventative, then there is scope for greater learning about B during Stage 2. This idea could be tested further by conducting another version of the experiment with a second group, in which B- is explicitly trained as inhibitory (i.e. preventative). The theory protection account predicts that greater learning about B compared to A should not be

observed in this second group, resulting in a between-group difference. Conversely, Pearce and Hall's model predicts no difference between the groups because all cues should decline in associability to an equal extent, since all trials have predictable outcomes.

Despite the two influential attentional models discussed above (Mackintosh, 1975 and Pearce & Hall, 1980) failing to account for the full set of experiments in Chapters 2 and 3, this does not mean that attention may not play some role in explaining the current results. There are more recent attentional accounts worthy of future investigation, such as EXIT (Kruschke, 2001), Locally Bayesian Learning (Kruschke, 2006), or the hybrid model proposed by Le Pelley (2004). For example, the two models proposed by Kruschke retain the essence of Mackintosh's account (i.e. a non-recurrent associative model in which there is re-allocation of attention to the best predictors) but differ in other respects. Thus, the failure of Mackintosh (1975) and Pearce and Hall (1980) to accommodate the results of Experiments 6 and 7 does not necessarily imply that other models would also fail. More broadly, there is no obvious theoretical reason why theory protection and attention should be mutually exclusive processes when attempting to construct a formal model. For example, the attention participants pay to cues may vary depending on how confident participants are about the causal status of those cues.

While discussing theory protection in the context of the present experiments, the predominant emphasis has been on participants protecting their theory about A during Stage 2. However, as briefly suggested in the previous section, participants should still hold a theory about B. However, this theory would be a more flexible one that encompasses a range of causal statuses. For example, in Experiments 4 and 5,

participants will know that cue B could either be causal, neutral or preventative before the first presentation of feedback in Stage 2. Therefore, learning that B is preventative, following Stage 2 feedback, allows participants to protect an aspect of this theory about B with relative confidence. This suggestion might at first seem tangential, but it illustrates an important point about how theory protection might operate more broadly. Learning that B is preventative of the outcome during Stage 2, does not contradict anything already known about B from Stage 1. For example, if B was instead trained as a blocked cue during Stage 1, participants should be unconfident about whether it is causal or neutral, but they should learn that it is not preventative of the outcome (because the outcome still occurs following the presentation of compounds containing a blocked cue). In such a design, participants would be unable to protect their existing theories about both A and B during AB- trials, since their existing knowledge about these cues would be inconsistent with the absence of the outcome. In Experiments 6 and 7, the data suggest that participants (prior to Stage 2) hold a theory that B is not a cause, but might be neutral or inhibitory. Therefore, participants can protect their theory that B is not a cause during Stage 2, while acquiring a more constrained theory that it is preventative. Again, learning that B is preventative during Stage 2 matches what they already know about B from Stage 1. Therefore, being unconfident about the status of a cue should only facilitate learning, if the range of causal statuses that cue could hold matches the subsequently experienced outcome.

This concept of matched (i.e. consistent) statuses and outcomes in theory protection is important for two reasons. Firstly, as already outlined above, as it raises the question of what might happen when an outcome is not consistent with the existing theories held about each cue in a compound. To provide another example, if a previously-trained

neutral cue and an inhibitor were presented in a compound that was followed by the outcome, the occurrence of this outcome would not be consistent with the existing theories held about either cue. When participants are unable to protect their existing theories, it is possible that learning might more closely resemble that predicted by prediction error accounts (e.g. Rescorla, 2001). This idea has yet to be tested, and seems a worthwhile avenue for future research. Secondly, as outlined in the Chapter 2 General Discussion (2.5), theory protection should still operate in instances where there is no difference in participants' confidence about the causal status of two cues being trained in compound (e.g. Le Pelley & McLaren, 2001). In such circumstances, participants should learn more about a cue, if its current causal status matches the outcome it is subsequently trained in compound with. The concept of matched statuses and outcomes in theory protection is discussed in further detail in Chapter 5.

The experimental scenario may also be important in influencing the type of theories participants form about cues. Theory protection in the Chapter 3 experiments relied on participants believing that B might be preventative, prior to receiving feedback in Stage 2. If participants do not believe that B can be an inhibitor then the opposite result should be observed, since participants would have no option but to revise their belief that A is causal. One way in which this idea could be tested is by varying the instructions provided to participants to manipulate their beliefs about cues. For example, participants could be presented with an experimental scenario in which cues are unlikely to be preventative. Haselgrove and Evans (2010) implemented a similar design to Experiments 6 and 7, using a scenario in which the cues were different foods and the outcome was a stomach ache. There were also some minor differences to their design, such as two additional compounds in the initial training stage. One of these resulted in

stomach ache, while the other resulted in no stomach ache. There was also another compound alongside AB- in the subsequent training stage, which resulted in no stomach ache. Additionally, their experiment compared two groups of participants, with high versus low scores on a measure of schizotypy. Unlike chemicals, foods are not generally regarded as something that can become preventative of an allergic reaction (Zaksaite & Jones, 2019). Consequently, participants should simply learn that B is neutral during the initial training stage. During subsequent AB- training, participants would be unable to attribute the absence of the outcome to B, and should therefore be forced to revise their theory that A is a cause of the outcome, perhaps in a manner consistent with individual prediction error. The results of Haselgrove and Evans (2010) support this view, as a compound test revealed more learning about A than B. However, it should be noted that this difference was significant in the high schizotypy group, but not the low schizotypy group. It would be useful to simplify this experiment, by dropping the schizotypy element, in order focus specifically on the effect of using food cues, as opposed to chemicals. This idea could be tested more directly in a between-groups experiment, in which the scenario and instructions are varied so that inhibition is encouraged in one group, but not the other.

Similarly to Chapter 2, the Test stage data reported in this chapter provided plots of inter-subject variability. As before, these plots suggest that a substantial minority of participants may be behaving in a way consistent with prediction error accounts (i.e. they individually have an AD-BC difference that is zero or negative). Again, future research might examine this issue in more detail, to look for stable individual differences in how associations are formed.

As a final point, the addition of the Probe Test and Forced Choice highlights an interesting issue concerning the best method of assessing predictive learning in humans. Typically, human predictive learning experiments (e.g. Spicer at al., 2019; Jones et al., 2019) ask participants to indicate the likelihood of the outcome, as a way of measuring learning. Our Probe Test instead asked participants specifically what they thought each cues does. This kind of approach might be a better test, in the case of predictive learning tasks more broadly, as it is sensitive to differences in causal status that outcome likelihood measurements are not. In the case of Experiment 7, the Probe Test allowed for differences between neutral cues and preventative cues to be demonstrated with individual ratings; something not possible when measuring outcome likelihood. More broadly, if cues can acquire negative associative strength, then it seems sensible for testing scales to incorporate a negative dimension as a standard feature.

4.1: Representing Uncertainty

Chapters 2 and 3 provide experimental evidence of theory protection playing an important role in human associative learning. Differences in participants' confidence about the causal status of cues was suggested as a way in which theory protection might operate. Specifically, if participants are less confident, then they should hold a weaker theory. The focus of this chapter is different, although there are some related themes. Firstly, there is a focus on individual and overall prediction error, cue competition, and the redundancy effect. These learning phenomena are formally investigated in this chapter, using computational model fitting. Secondly, there is a proposed method of formally implementing participants' lack of confidence (i.e. uncertainty) within the confines of an established model of learning; the Rescorla and Wagner (1972) model (N.B. popularly referred to as the Rescorla-Wagner model). This mathematical representation of uncertainty is specific to novel cues (i.e. cues not previously encountered in learning scenarios). This chapter does not focus specifically on theory protection. However, Chapter 5 suggests future avenues for research, with a particular focus on how a formal model of theory protection might be developed.

When formally modelling animal learning, the associative strengths of previously non-encountered cues typically start at zero (e.g. Rescorla & Wagner, 1972). When equivalently modelling human learning, it has been assumed that associative strengths should similarly start at zero. It has also been assumed that the Rescorla and Wagner (1972) model adequately accommodates the results of simple blocking experiments in humans (e.g. Dickinson, Shanks, & Evenden, 1984), whilst the Bush and Mosteller (1951) model cannot. The simulations presented in this chapter demonstrate that both of

these assumptions are wrong. The Rescorla-Wagner model, as usually applied, fits the results of a simple blocking experiment no better than Bush & Mosteller's (1951) model. However, if the Rescorla-Wagner model is modified, so that the initial associative strengths of cues can be an intermediate value, rather than zero, then the model does indeed provide a better account of simple blocking data than the (equivalently modified) Bush and Mosteller model. This modification also allows the Rescorla-Wagner model to account for a redundancy effect (e.g. Uengoer et al., 2013) experiment; something that the unmodified model is not able to do. Furthermore, the modified Rescorla-Wagner model can accommodate the effect of varying the proportion of trials on which the outcome occurs (i.e. the base rate) on the redundancy effect (Jones et al., 2019). Interestingly, the initial associative strength of cues varies in line with the outcome base rate. The key proposal of this chapter is that this modification provides a simple way of mathematically representing uncertainty about the causal status of novel cues within the confines of the Rescorla-Wagner model. The theoretical implications of this modification are discussed. Free and open resources to support formal modelling in associative learning are introduced.

As discussed in Chapter 2, blocking (Kamin, 1969) is the most widely known type of cue competition in associative learning. Learning about blocked cues is apparently restricted by the simultaneous presence of another cue that has also been trained separately (e.g. cue X from an A+ AX+ design). Blocking has long been of interest to associative learning researchers. Early models of associative learning were unable to account for it. In particular, the associative learning model developed by Bush and Mosteller (1951) uses an individual prediction error, which means that any change in the strength of an association between a cue and an outcome is governed by the size of

the error between the outcome that occurs and the outcome predicted by that cue alone. For example, if you predict that a certain type of food is safe to eat, but you subsequently suffer an allergic reaction after eating that food, then there will be a large prediction error. Once you have learned that this type of food is not safe to eat, there will be no error, and learning will be at asymptote. According to the Bush and Mosteller model, learning updates algorithmically as follows:

$$\Delta V_x = \alpha_x L(\lambda - V_x) \quad (1)$$

In Equation 1, associative strength is denoted by V, where ΔV_x is the change in associative strength for cue X, and V_x is the current associative strength of cue X. The cue salience is represented by α and the outcome learning rate is represented by L. The asymptote of learning is represented by λ. The individual prediction error term means that the model cannot account for blocking, as this effect is driven by the interference of simultaneously encountered cues. To overcome this, Rescorla and Wagner (1972) proposed their model with an overall prediction error, in which any change in associative strength is governed by the error between the outcome that occurs and the outcome predicted by all simultaneously present cues. The equation is as follows:

$$\Delta V_x = \alpha_x L(\lambda - \Sigma V) \quad (2)$$

The only change is that ΣV has been incorporated as the overall associative strength of all simultaneously-present cues. Such cues compete for the available associative strength, allowing the model to account for blocking. This is because A takes up all the available associative strength on the A+ trials, meaning that there is no available

associative strength for X to acquire on the AX+ trials (assuming learning is at asymptote). In other words, X is left with nothing to account for, as outcome on the AX+ trials is fully predicted by A.

Dickinson et al. (1984) provided one of the first demonstrations of blocking in humans. As stated, the Rescorla-Wagner model is widely assumed to provide an adequate explanation of such human blocking experiments, whilst also providing a better account than Bush and Mosteller's model. The design Dickinson et al. (1984) employed was as follows: training trials containing cue A were presented in the first stage of the experiment, while the trials containing the AX+ compound were presented in a subsequent stage (forward blocking). In the first simulation reported in this chapter, a model fitting procedure was conducted on the full data of a simple forward cue competition blocking experiment, using both the Rescorla-Wagner, and Bush and Mosteller models. It is worth emphasising that the ability of the models to account for the full set of experimental test cues was explored, rather than just the blocked cue and the control cue. In other words, a sufficiently adequate model should be able to account for a full experiment, rather than just accounting for an effect, which is only an extract of an experiment. Neither model provided an adequate account of the full test data, with the Rescorla-Wagner model's overall fit to the data being no better than Bush and Mosteller's model.

4.2: Model fitting 1: blocking

There are several simple human forward cue competition experiments reported in the literature (e.g. Dickinson et al., 1984; Miller, 1996; Mitchell & Lovibond, 2002), but none of these datasets appear to have been made openly accessible. In order to provide an openly accessible set of simple blocking data, a standard forward cue competition experiment was conducted. The design included blocking (A+ AX+), a common control (B- BY+), and filler cues intended to balance the number of cue types causing either stomach ache or no stomach ache (C- CD-). The full details of the experiment are included in the next section of this chapter, followed by details about the model fitting procedure.

4.2.2: Blocking experimental details (Experiment 8)

Consistent with a number of previous human learning experiments (e.g. Uengoer et al., 2013; Jones et al., 2019), this experiment used a food allergy scenario. Participants were provided with a fictional situation, in which they played the role of a medical doctor, trying to ascertain which foods cause a stomach ache in a patient. During the training trials, participants were presented with a series of different foods, either singly or in pairs. They were asked to predict whether or not the fictional patient would experience a stomach ache after eating these foods. After participants had made their prediction, they were provided with feedback as to whether or not a stomach ache occurred. Once participants had completed training, they were tested by being asked to make a likelihood rating indicating how likely they thought a stomach ache would be after the patient ate specific foods.

Table 4. The design Experiment 8

Stage 1	Stage 2	Test
A+	AX+	A D
B-	BY+	B X
C-	CD-	C Y

The experimental design is shown in Table 4. The Stage 1 training consisted of three single cues, one of which was followed by the outcome (A+), and two of which were followed by the absence of the outcome (B- C-). The subsequent Stage 2 training consisted of three compounds, each of which contained one cue from Stage 1. Two of these compounds were followed by the outcome (AX+ BY+), and one was followed by the absence of the outcome (CD-). As outlined in the chapter introduction, X was added as the blocked cue, while Y was added as a control (known as a feature positive

control). If X is judged to be a less likely cause of the outcome at test than Y, then this provides evidence of blocking. The individual cue C- and the compound CD- were added as filler cues, so that an equal number of trial types were followed by either stomach ache or no stomach ache across the two training stages.

Method

Participants

Forty-one psychology students from the University of Plymouth participated in this experiment, in return for course credit (31 female, 10 male; mean age = 19.59, SD = 1.41). This sample size has adequate power to detect medium-sized within-subjects effects (88% power at d = 0.5). People who had previously taken part in similar experiments were excluded from this study, to ensure that participants were naive to the purpose of the experiment.

Materials

Participants were all tested in the same lab at the University of Plymouth. The experiment was conducted using Viglen Genie desktop computers, running the Windows 10 operating system. The computers all used 22-inch Phillips LED displays, with participants at a typical distance (of approximately 40-80 cm) from the screen. The experiment was designed and executed in the Psychopy desktop application version 1.85.2 (Peirce, 2007). Participants made their responses by pressing keys on a standard UK computer keyboard during the training stages, and by using mouse clicks during the test stage. The six individual cue types were represented on screen as photographs of fruits: apple, banana, kiwi, orange, pear, and plum. All the fruits were presented within a white square. The dimensions of each cue (including the white square) were 300 x 300 pixels, with a screen resolution of 1920 x 1080 pixels. For each participant, the foods were randomly assigned to each cue (A, B, C, D, X, and Y). The two outcomes, ‘stomach ache’ and ‘no stomach ache’, were represented by text on screen and a

photograph of a man clutching his stomach, or a man giving a ‘thumbs up’, respectively. The outcome images were presented within a white rectangle. The dimensions of the outcome images (including the white rectangle) were 291 x 332 pixels. All experimental text, including instructions, was white. A black background was used throughout the experiment. Study information sheets, consent forms and debrief forms were all printed on paper.

Design

The experiment used a within-subjects design, as outlined in Table 4. During Stage 1, participants were presented with twelve blocks of training. The three trial types (A+, B-, C-) were presented once within each block, and in a random order. During Stage 2, participants were presented with six blocks of training. The three trial types (AX+, BY+, CD-) were presented once within each block, and in a random order. During the Test stage, participants were presented with ten blocks of test cues. The purpose of this extended Test stage, was to allow these data to potentially be used for model fitting at the level of individual participants (N.B. this is not the intention of the current project). The six trial types (A, B, C, D, X, Y) were presented once within each block, and in a random order.

Procedure

Participants were required to read an information sheet and sign a consent form prior to participating in the experiment. The experimental instructions were presented on the

screen at the start of the experiment. They were adapted from Uengoer et al. (2013) and were as follows:

This study is concerned with the way in which people learn about relationships between events. In the present case, you should learn whether the consumption of certain foods leads to stomach ache or not.

Imagine that you are a medical doctor. One of your patients often suffers from a stomach ache after eating. To identify which foods they react to, the patient eats specific foods and observes whether a stomach ache occurs or not. The results of these tests are shown to you on the screen one after the other.

You will always be told what your patient has eaten. Sometimes, they have only consumed a single kind of food and on other times they have consumed two different foods. Please look at the foods carefully.

You will then be asked to predict whether the patient suffers from stomach ache. For this prediction, please click on the appropriate response button. After you have made your prediction, you will be informed whether your patient actually suffered from stomach ache. Use this feedback to find out what foods cause a stomach ache in your patient.

At first you will have to guess the outcome because you do not know anything about your patient. But eventually you will learn which foods lead to stomach ache in this patient and you will be able to make correct predictions.

For all of your answers, accuracy rather than speed is essential. Please do not take any notes during the experiment. If you have any more questions, please ask them now.

If you do not have any questions, please start the experiment by pressing the space bar.

For each trial during the training stages, the cues were presented visually on either the left- or right-hand side of the screen. Since only one image was presented on each trial in Stage 1, the opposite side of the screen contained a blank space, matching the black background. The cues were randomly assigned to either the left- or right-hand position on each trial. Text at the top of the screen stated that 'The patient eats the following:', with the stimuli presented below this. Underneath the stimuli, further text stated 'Which outcome do you expect? Please use your keyboard to respond'. Participants were instructed to respond by pressing the appropriate key on their keyboard; Z for 'No Stomach Ache' and M for 'Stomach Ache'. After participants made their response, the feedback for that trial was shown. The feedback screen consisted of the appropriate outcome image along with its accompanying text, indicating either 'Stomach Ache' or 'No Stomach Ache'. The feedback was shown on screen for two seconds, after which the next trial began.

Following the completion of Stage 1, Stage 2 started with no trial break, so that this was a seamless continuation of the training from the perspective of participants. As in Stage 1, the cues were randomised on each trial to appear on either the left- or right-hand side of the screen. Since two cues were presented on each trial, there was no need for a blank space to be presented during this stage. The on-screen text and responding via the

keyboard was the same as in Stage 1. The process for displaying the trial feedback was also the same.

After the completion of Stage 2, a further instruction screen was shown before commencement of the Test stage:

Now, your task is to judge the probability with which specific foods cause stomach ache in your patient. Single foods will be shown to you on the screen.

In this part of the experiment, you will receive no feedback about the actual reaction of the patient. Use the information that you have collected so far, to make your rating.

Press the space bar to continue.

For each trial during the Test stage, the cues were presented on either the left- or right-hand side of the screen. When only one image was presented, the opposite side of the screen again contained a blank space. The cues were randomly assigned to either the left or right position on each trial. As before, text at the top of the screen stated that 'The patient eats the following:', with the stimuli presented below this. Underneath the stimuli, further text stated 'How likely are they to suffer a stomach ache? (0 = Very Unlikely; 10 = Very Likely)'. Participants were instructed to respond by clicking on an eleven-point rating scale using their mouse pointer, to indicate how likely they thought the occurrence of a stomach ache would be. The rating scale was located in the lower part of the screen, with the 11-point scale running from left to right, in ascending numerical order. After participants made their response, a black screen appeared for 0.4

secs, after which the next trial was presented. Following the completion of the experiment, participants were provided with a debrief form.

Analysis

The data were processed and analysed using R (R Core Team, 2018). The difference between cues X and Y (i.e. to test for blocking) was assessed using a paired-samples t-test. The alpha level was set to $p < .05$. A Bayesian t-test was also conducted, using the procedure recommended by Dienes (2011) and implemented as R code by Baguley and Kaye (2010). Due to methodological differences among previous demonstrations of forward cue-competition in humans, it was difficult to choose a specific experiment on which to base a prior. Therefore, a uniform distribution was specified, with a lower limit of -10 and an upper limit of 10 (in terms of the mean difference between ratings), because this is the largest difference between ratings permitted by an 11-point scale. In keeping with accepted conventions (e.g. Jeffreys, 1961), a Bayes factor of over three was set as the level providing evidence for a difference, while a Bayes factor of less than one third was set as the level providing evidence for no difference. A value between these levels was accepted as being inconclusive.

Results and Discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/sa8ux/. The descriptive statistics for the Experiment 8 training stages are shown in Figure 16. The Stage 1 data indicate that participants learned sufficiently about the three different cues by the time this stage was complete. Similarly, the Stage 2 data indicate that participants learned sufficiently about the three different compounds by the time training was complete.

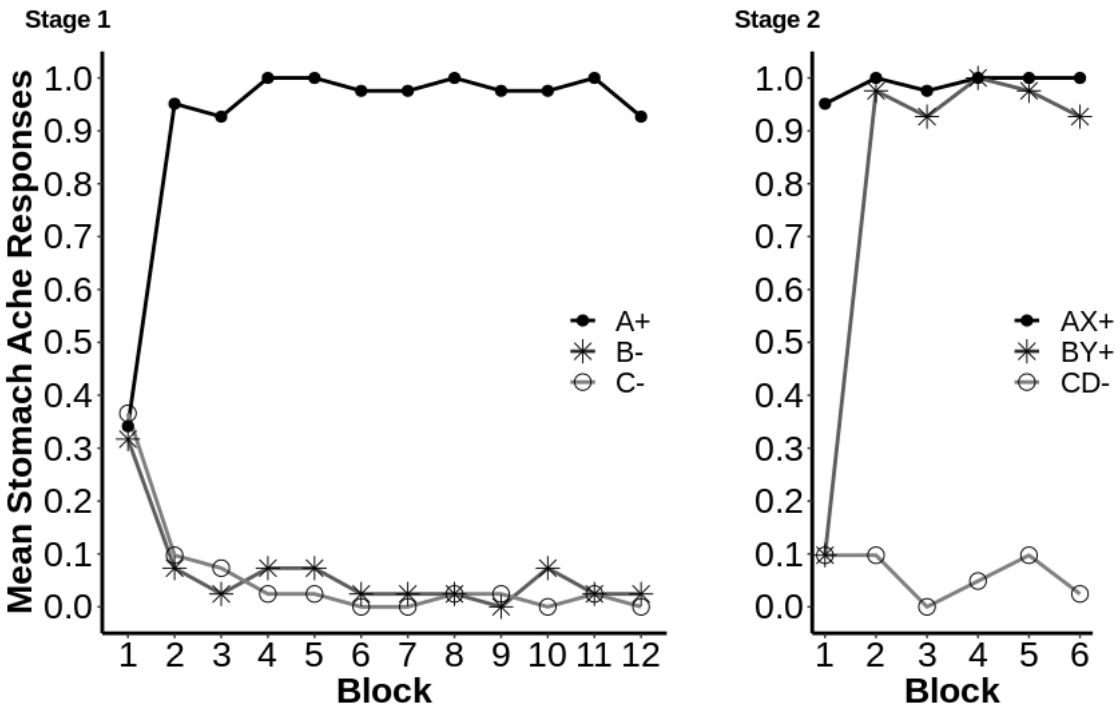

Figure 16. Experiment 8 Stage 1 and 2 data.

Descriptive statistics for the Experiment 1 Test Stage are shown in Figures 18 and 19, along with the model fitting data (see next section). Mean ratings for X were significantly lower than for Y; $t(40) = 6.77$, $p < .001$, $BF = 5.66 \times 10^8$, $d = 1.06$, indicating a successful demonstration of blocking. Figure 17 shows inter-subject variability on the key test difference. As would be expected from the mean test ratings,

most individual participants rated X lower than Y, although the variability did extend to some participants assigning a lower rating to Y.

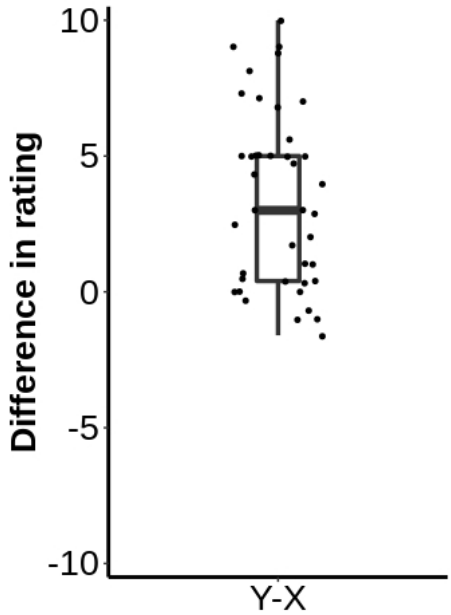

*Figure 17. I*nter-subject variability on the key Y-X difference using a method comparable to Hintze & Nelson (1998). Each dot is one participant, with jitter applied for readability. The boxplot shows the median and interquartile range.

4.2.3: Blocking model fitting details

A model fitting process was conducted on the data for all test stage cues. The model fitting was conducted using standard implementations of the Bush and Mosteller (1951) model and the Rescorla and Wagner (1972) model. The fitting code for this specific simulation is available at https://osf.io/sa8ux/. Following standard practice, the procedure used gradient-descent optimisation to find the best fitting parameters. This process involves exploring the parameter space, to minimise the difference (represented by the sum of squared errors) between the observed and predicted mean test ratings. It was necessary for each model to generate ratings consistent with the outcome likelihood scale (between 0 and 10) used at test, rather than just output associative strengths (typically between 0 and 1). As outlined in Chapter 1, it is generally accepted that there is not a 1:1 linear relationship between associative strengths and responding (e.g. Gluck & Bower, 1988), although as one increases so should the other. A standard solution to this problem is to use a logistic function to map associative strengths onto responses. The equation below is the logistic function suggested by Gluck and Bower:

$$P_x = 10 \frac{1}{1+e^{-\Theta(V_x-\beta)}} \quad (3)$$

P_x denotes the simulated likelihood rating for cue X, while V_x denotes the associative strength. β (Beta) denotes the bias parameter for the output associative strength that will result in a rating of 5 (i.e. the middle of the scale). Θ (Theta) is a scaling parameter, where higher values mean that the function relating activation to rating becomes less linear and more logistic. At high values it produces a step function. It should be noted that this function was originally proposed as a way of mapping associative strengths

onto choice probabilities in category learning. However, predictive learning tasks, such of those reported in this thesis rely on participants making a probabilistic choice, in terms of whether an outcome will occur or not. As the probability (i.e. likelihood) ratings participants make are an indication of underlying associative strength, this function provides an appropriate way of mapping this relationship. Indeed, Rescorla (2001) describes such a logistic (i.e. sigmoidal) function for mapping associative strength onto responding, although does not suggest a formal implementation. As outlined in the introduction to this chapter, both models require cue salience and learning rate as their two standard parameters. Both of these parameters can have a value between 0 and 1. Since these values are multiplied in the learning algorithm, the resulting product is necessarily a value between 0 and 1. For simplicity, these parameters were collapsed into a single learning rate parameter (LR). Therefore, the parameter space being explored consisted of LR, Beta and Theta. The value of each parameter was the same for all cues, since the counterbalancing of stimuli in the experimental design meant there was no theoretical basis for expecting these values to differ between cues.

Results from standard model implementations

The code for producing the model fitting outputs is available at https://osf.io/sa8ux/. An optimisation process was conducted on the blocking data, in which the parameter space for the two models was explored (using the *optim* function of R) to find the best fitting parameters. The difference between the predicted and observed test ratings, as represented by sum of squared errors (SSE) was minimised in order to find the closest fit. The training that each experimental participant received was represented as a training matrix, which was passed to each model, so that the simulated participants received the 'same training' as the experimental participants. The model implementations are part of the *catlearn* (Wills, Dome, Edmunds, Honke, Inkster, Schlegelmilch, & Spicer, 2019) package in R. The Rescorla and Wagner (1972) implementation is called *slpRW*, and the Bush and Mosteller (1951) implementation is called *slpBM*. The function for defining the relationship between associative strength and responding is also available in *catlearn* and is called *act2probrat* (this is an implementation of Equation 3). The best fitting parameters for the standard (unmodified) implementations of the Rescorla-Wagner, and Bush and Mosteller models are reported in Table 5. The learning rate is reported as LR, whilst Beta and Theta are the parameters that define the relationship between associative strength and responding. The sum of squared errors (SSE) and the mean error (i.e. for each cue) are reported, along with the adequacy of fit (R^2). Both models provided an equally inadequate account of the blocking test data.

Table 5. Output of unmodified model fitting on blocking data

Model	**LR**	**Beta**	**Theta**	**SSE**	**Mean Error**	**R^2**
BM	0.04	0.21	13.41	0.24	0.08	0.71
RW	0.01	0.04	75.81	0.24	0.08	0.71

The best fitting model should produce the smallest error between the predicted and observed test ratings. This was assessed using the mean error for each of the six test cues. The adequacy of fit for each model was additionally assessed with the R^2 of the predicted versus observed ratings (this is a standard correlation co-efficient, and hence a higher value means a better adequacy of fit). Both the Bush and Mosteller model, and the Rescorla-Wagner model produced an equivalent mean error and R^2. To put the R^2 value in context, the best formal models of category learning produce R^2 values exceeding .95 for standard results in the field, with models that are clearly and ordinally wrong still sometimes producing R^2 values exceeding 85% (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). On this basis, both the Rescorla-Wagner, and Bush and Mosteller models provide rather poor accounts of this basic blocking experiment.

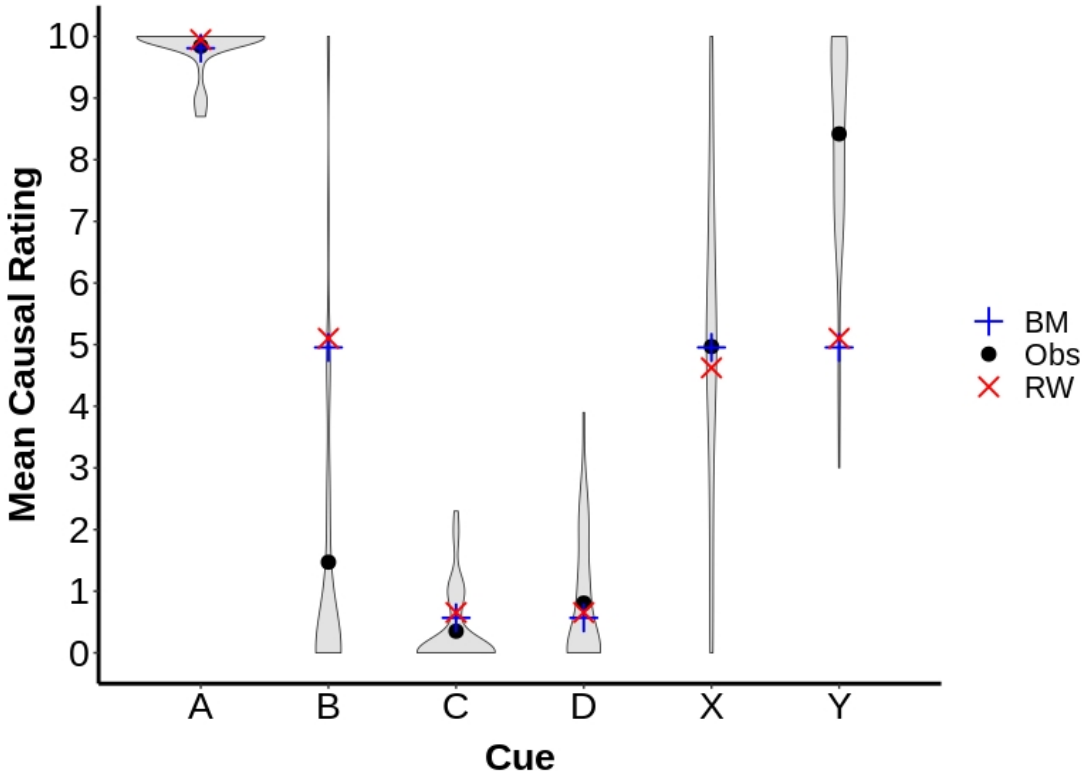

Figure 18. Predicted versus observed Test stage ratings for unmodified Rescorla-Wagner (RW) model, and Bush & Mosteller (BM) model, against observed data (Obs), following A+/AX+ B-/BY+ C-/CD- training. The violin plot represents the distribution of the observed data using a method comparable to Hintze & Nelson (1998). The distribution of the predicted data is not represented, as it was negligible.

Figure 18 shows the predicted versus observed test cue ratings for both models. The unmodified Rescorla-Wagner model provides no better an account of the blocking data than the Bush and Mosteller model. The Bush and Mosteller model cannot capture these data, because B, X and Y caused stomach ache on an equal number of Stage 2 training trials, so learning should be equal according to individual prediction error. However, the Rescorla-Wagner model also predicts equal learning about B and Y, since they would equally account for the occurrence of stomach ache during the BY+ trials. The Rescorla-Wagner model fitted X, but at the cost of getting B and Y wrong in opposing directions. The difference between X and Y was under-predicted. This could either be interpreted as an under-prediction of the blocking effect itself, or as an inability of the

model to capture protection from overshadowing (i.e. the effect of the B- trials on responding to Y).

4.2.4: Modifying the Rescorla-Wagner model

The Rescorla and Wagner (1972) model was originally developed as an account of non-human animal learning. In that context, it makes sense for the associative strength of cues to start at zero, because animals such as rats would not have learned a response to cues not previously encountered. However, in a human predictive learning context, it seems unlikely that a novel cue would have an associative strength of zero. This is because participants would be learning whether or not such cues are the cause of an outcome, and an associative strength of zero should result in the production of low causal ratings. A more intuitive response would be for participants to provide novel cues with an intermediate rating (for example 5 on a scale running from 0-10), reflecting their uncertainty about the causal status of those cues. This is supported by the results of Experiment 4, in which the novel cue at test (D) was assigned an intermediate rating of 4.85 on the 0-10 likelihood scale. The model fitting procedure was therefore conducted on the blocking data a second time, using modified versions of the models, in which the initial associative strength of cues was an additional parameter for optimisation. As with the other parameters, the initial associative strength was the same for all cues, since there was no theoretical basis for expecting any differences. The fitting code for this simulation is available at https://osf.io/sa8ux/.

Results from modified model implementations

The best fitting parameters for these modified implementations of the Rescorla and Wagner (1972), and Bush and Mosteller (1951) models are reported in Table 6. The best fitting initial associative strength (Init Assoc) is reported as the additional parameter. The predicted versus observed ratings for these best fitting parameters are reported in Figure 19.

Table 6. Output of modified model fitting on blocking data

Model	**LR**	**Init Assoc**	**Beta**	**Theta**	**SSE**	**Mean Error**	**R^2**
BM (modified)	0.02	0.45	0.49	30.52	0.06	0.04	0.93
RW (modified)	0.04	0.43	0.41	19.31	0.00	0.01	1.00

If the Rescorla-Wagner model is modified so that the starting associative strength can be something other than zero, then it accounts for the results of the blocking experiment better than the equivalently modified Bush and Mosteller (1951) model, as indicated by the mean error and R^2 values. The Rescorla-Wagner model produces less error and has a better adequacy of fit. Whilst the modified Bush and Mosteller model provided a better fit than the unmodified version, the lack of an overall error term does not allow it to predict blocking.

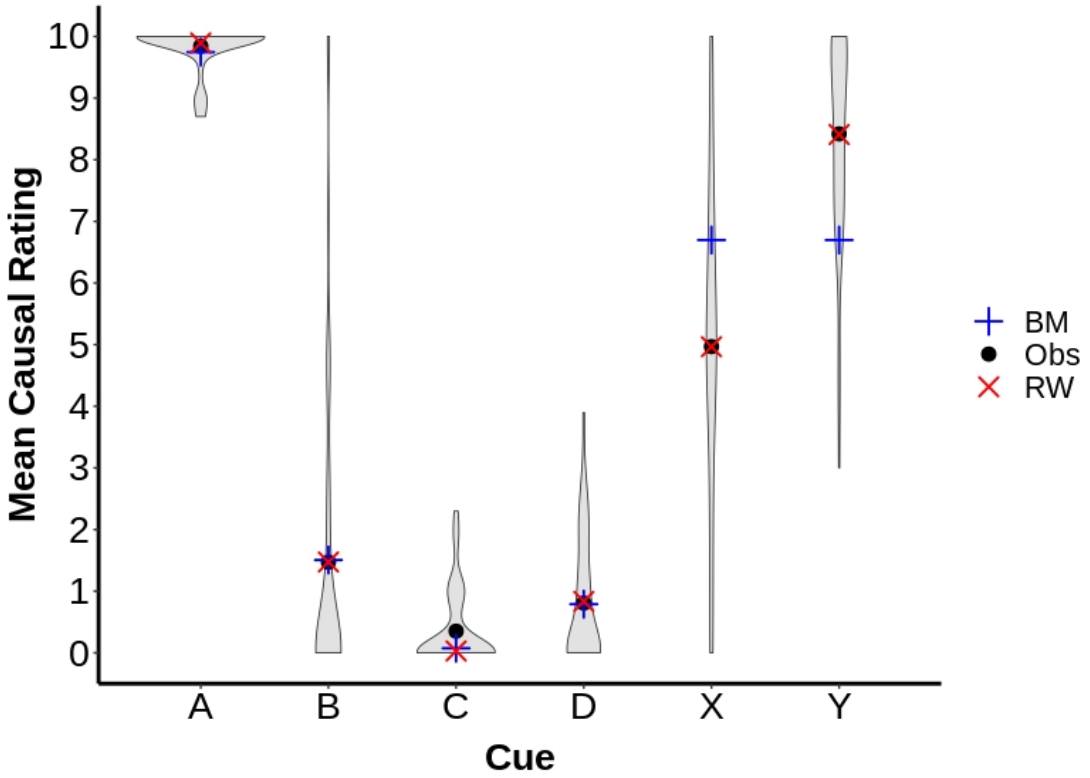

Figure 19. Predicted versus observed Test stage ratings for modified Rescorla-Wagner (RW) model, and Bush & Mosteller (BM) model, against observed data (Obs), following A+/AX+ B-/BY+ C-/CD- training. The violin plot represents the distribution of the observed data.

As predicted, the best fitting initial associative strengths were at an intermediate value for both models. This finding is consistent with the idea that participants should assign intermediate ratings to cues that have an unknown causal status.

4.3: Model fitting 2: redundancy effect

The next step was to investigate whether the modified Rescorla-Wagner model could adequately capture a further psychological phenomenon that the unmodified model cannot; the redundancy effect (e.g. Jones & Zaksaite, 2018; Jones et al., 2019; Uengoer, Lotz, & Pearce, 2013; Uengoer, Dwyer, Koenig, & Pearce, 2019). Recall from Chapter 2 that the training stage of a redundancy effect design incorporates blocking (A+ AX+) and a simple discrimination (BY+ CY-). Cue Y is referred to as an uncorrelated cue, because it appears in both a causal compound and a non-causal compound. The redundancy effect is the observation of X being rated as a more likely cause of the outcome than Y. The blocked cue (X) is typically given intermediate causal ratings at test, while the uncorrelated cue (Y) is typically given low causal ratings.

Rather than using a previously published data set, a new redundancy effect experiment was conducted, in order to provide suitable data. The motivation for collecting new data was a need for a more diagnostic test stage than simply asking participants to provide likelihood ratings for the five single cues (A, B, C, X, Y). In addition to asking participants to provide likelihood ratings for the five single cues at test, they were also asked to provide ratings for each of the ten possible compound cue pairs (that can be produced using these five individual cues). Importantly, the training participants received was equivalent to the training used in previous redundancy effect demonstrations (A+ AX+ CY+ CY-). However, having a wider set of cues in the test Stage meant that the two models could be ‘stretched’, by being required to fit a more complex set of test data. The fitting code for this simulation is available at https://osf.io/9vaby/.

4.3.2: Redundancy effect experimental details (Experiment 9)

This experiment used the same scenario as Experiment 8. The presentation of the cues and feedback was the same, as was the responding by the participants. The experimental design is shown in Table 7. The Stage 1 training consisted of a simple intermixed blocking design (A+ AX+), along with a simple intermixed discrimination (BY+ CY-). If X is judged as a more likely cause of the outcome than Y at test, then this provides evidence of the redundancy effect.

Table 7. The design of Experiment 9

Stage 1	Test
A+	A AB BX
AX+	B AC BY
BY+	C AX CX
CY-	X AY CY
	Y BC XY

Method

Participants

Forty psychology students from the University of Plymouth participated in this experiment, in return for course credit (33 female, 7 male; mean age = 20.95, SD = 5.30). This sample size has adequate power to detect medium-sized within-subjects effects (87% power at d = 0.5). People who had previously taken part in similar experiments were excluded from this study, to ensure that participants were naive to the purpose of the experiment.

Materials

The Experiment 9 materials were the same as those used in Experiment 8, although a different version of Psychopy was used; 1.83.04 (Peirce, 2007). Additionally, a slightly different set of photographs (apple, banana, orange, pear, and strawberry) was used to represent the five individual cue types (A, B, C, X, and Y).

Design

The experiment used a within-subjects design, as outlined in Table 7. During Stage 1, participants were presented with four epochs of training. The four trial types (A+, AX+, BY+, CY-) were presented twice within each epoch. The reason why each trial type was presented twice was so that each of the two possible left-right cue configurations was counterbalanced. The trial types were presented in a random order. During the Test stage, participants were presented with six epochs of test cues. The purpose of this

extended Test stage, was to allow this data to potentially be used for model fitting at the level of individual participants (N.B. this is not the intention of the current project). The fifteen trial types (A, B, C, X, Y, AB, AC, AX, AY, BC, BX, BY, CX, CY and XY) were presented twice within each epoch (again to counterbalance the left-right cue configurations), and in a random order.

Procedure and analysis

The procedure and analysis for Experiment 9 was the same as for Experiment 8 apart from two minor changes. Firstly, the instructions before the Test Stage were adjusted to reflect the fact the there would be compound trials as well as single cue trials:

Now, your task is to judge the probability with which specific foods cause stomach ache in your patient. Both single foods and pairs of two different kinds of food will be shown to you on the screen.

In this part of the experiment, you will receive no feedback about the actual reaction of the patient. Use the information that you have collected so far, to make your rating.

Press space bar to continue the experiment

Secondly, the results of Uengoer et al. (2013) were selected for the specification of a plausible predicted effect size for the Bayesian t-test. The mean difference from the first test stage of their Experiment 2 was chosen, because participants received equivalent training and testing to the current experiment.

Results and Discussion

The trial-level raw data and analysis script for this experiment are available at https://osf.io/9vaby/. The descriptive statistics for the Experiment 9 training stage are shown in Figure 20. The data indicate that participants learned sufficiently about the four different trial types by the time Stage 1 was complete.

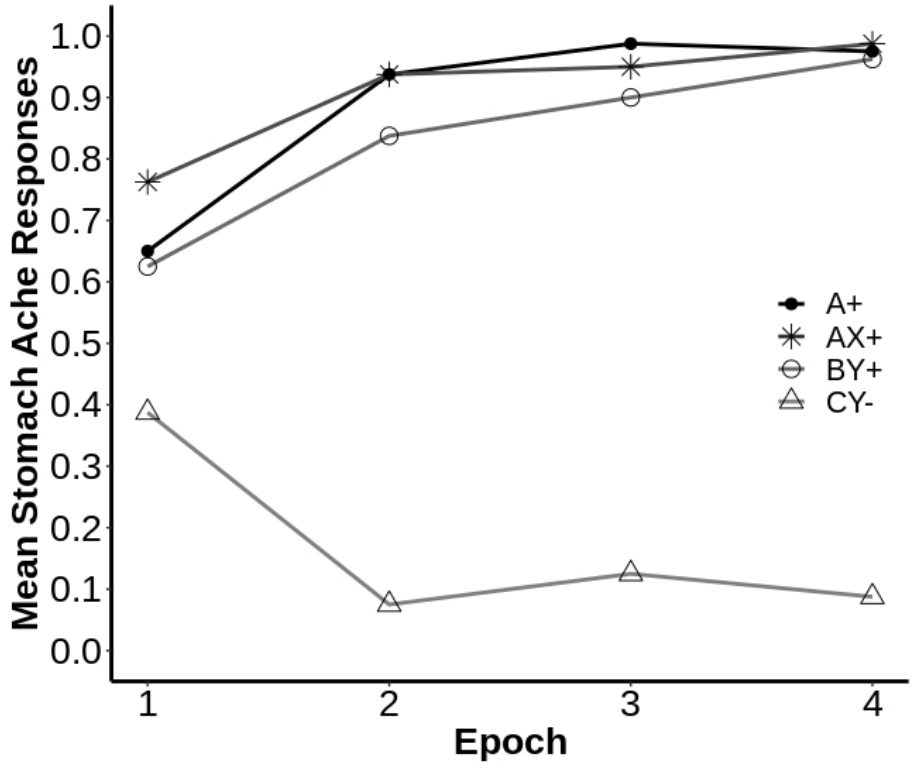

Figure 20. Experiment 9 training data

Descriptive statistics for the Experiment 9 Test Stage are shown in Figures 22 and 23, along with the model fitting data (see next section). Ratings for X were significantly higher than for Y; $t(39) = 2.69$, $p = .011$, $BF = 18.87$, $d = 0.43$, indicating a successful demonstration of the redundancy effect. Figure 21 shows inter-subject variability on the key test difference. Despite the mean difference between X and Y, the variability on the

X-Y difference extended to a number of participants assigning a higher rating to Y than they did to X.

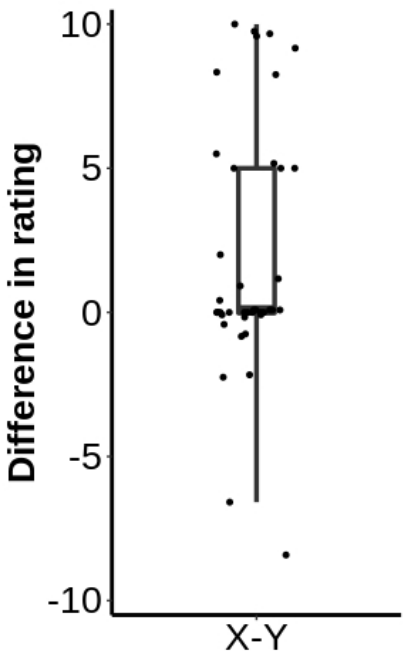

Figure 21. Inter-subject variability on the key X-Y difference.

4.3.3: Redundancy effect model fitting details

The code for producing the model fitting outputs is available at https://osf.io/9vaby/. The model fitting process conducted on the redundancy effect data used the same methodology as with the blocking data.

Results from standard model implementations

Table 8 shows the best fitting parameters, error and adequacy of fit for the unmodified implementations of both models. The Rescorla-Wagner model performed better in this instance, although the adequacy of fit is still lower than what would be expected from a well performing model, as outlined in the blocking section of this chapter. Issues with the fit for both models can be seen in the predicted versus observed test data.

Table 8. Output of unmodified model fitting on redundancy effect data

Model	**LR**	**Init Assoc**	**Beta**	**Theta**	**SSE**	**Mean Error**	**R^2**
BM (standard)	0.12	N/A	0.64	3.17	0.52	0.05	0.62
RW (standard)	0.45	N/A	0.34	3.09	0.24	0.03	0.83

Figure 22 shows the predicted versus observed test data for the unmodified Rescorla-Wagner, and Bush and Mosteller models. Crucially, the unmodified Rescorla-Wagner model is not able to capture the redundancy effect (X > Y), as anticipated. The Bush and Mosteller model is able to capture the redundancy effect itself, but does not provide an adequate account of the test data as a whole.

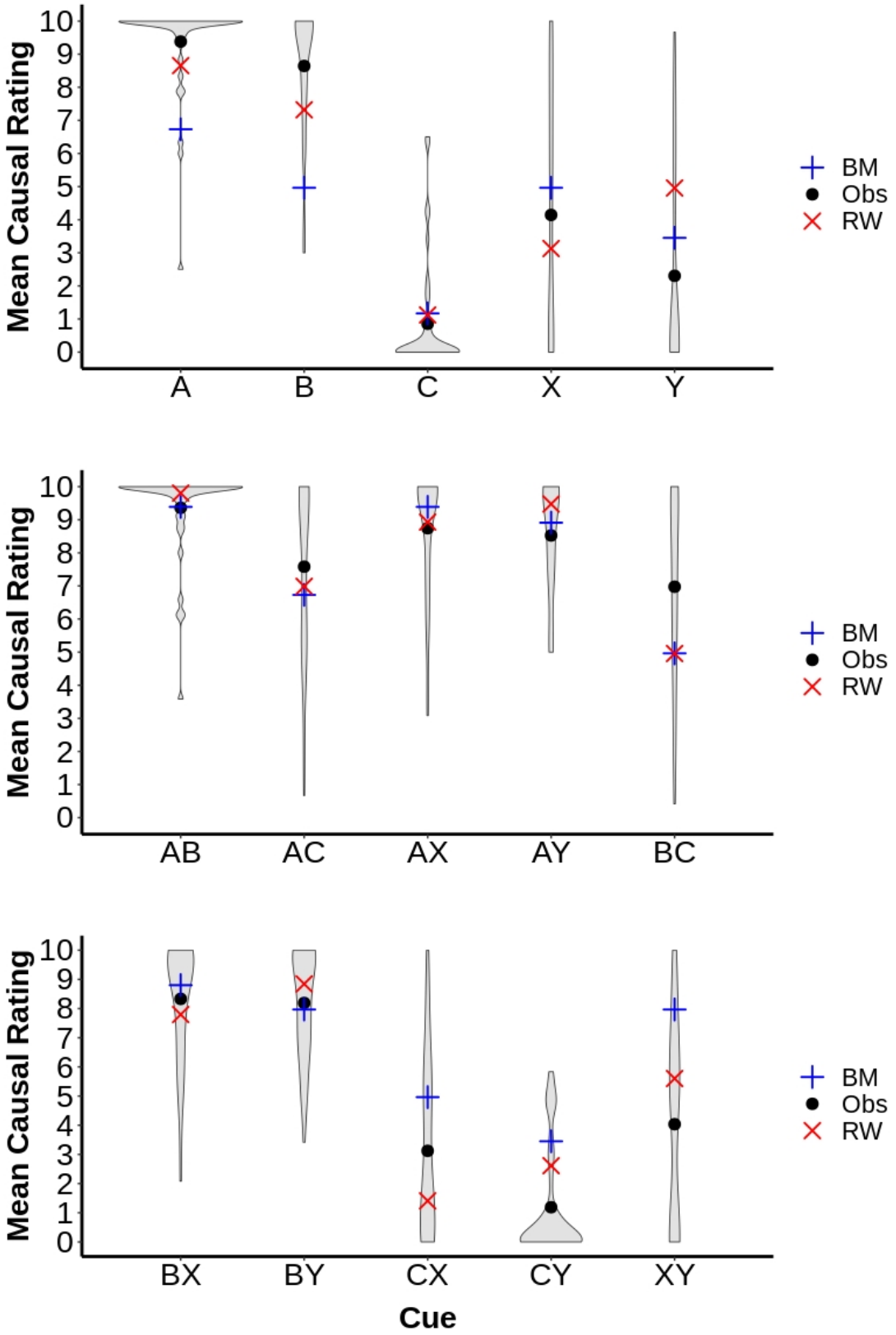

Figure 22. Predicted versus observed Test stage ratings for unmodified Rescorla-Wagner (RW) model, and Bush & Mosteller (BM) model, against observed data (Obs), following A+ AX+ BY+ CY- training. The violin plot represents the distribution of the observed data.

Results from modified model implementations

Table 9 shows shows the best fitting parameters, error and adequacy of fit for the modified implementations of both models. The modified implementation of the Rescorla-Wagner model produced less error and a better adequacy of fit than the unmodified implementation, while the modified implementation of the Bush and Mosteller model was no better than the unmodified implementation at capturing the experimental data. This indicates that the modified Rescorla-Wagner model is not performing better simply as a consequence of over fitting, from the inclusion of an additional free parameter for the starting associative strength.

Table 9. Output of modified model fitting on redundancy effect data

Model	**LR**	**Init Assoc**	**Beta**	**Theta**	**SSE**	**Mean Error**	**R^2**
BM (modified)	0.14	-0.34	0.57	3.03	0.51	0.05	0.63
RW (modified)	0.26	0.62	0.44	3.58	0.05	0.02	0.96

Importantly, these data indicate that the modified Rescorla-Wagner model is a good account of the dataset, while the modified Bush and Mosteller model provides a rather poor account, with no discernible improvement on the unmodified model. Figure 23 shows the predicted versus observed test data for the modified Rescorla-Wagner, and Bush and Mosteller, models. The modified Rescorla-Wagner model is able to capture the redundancy effect, although the size of the effect is somewhat underestimated. It is worth noting that the modified Bush and Mosteller model can also account for the redundancy effect itself, but not the full set of test data.

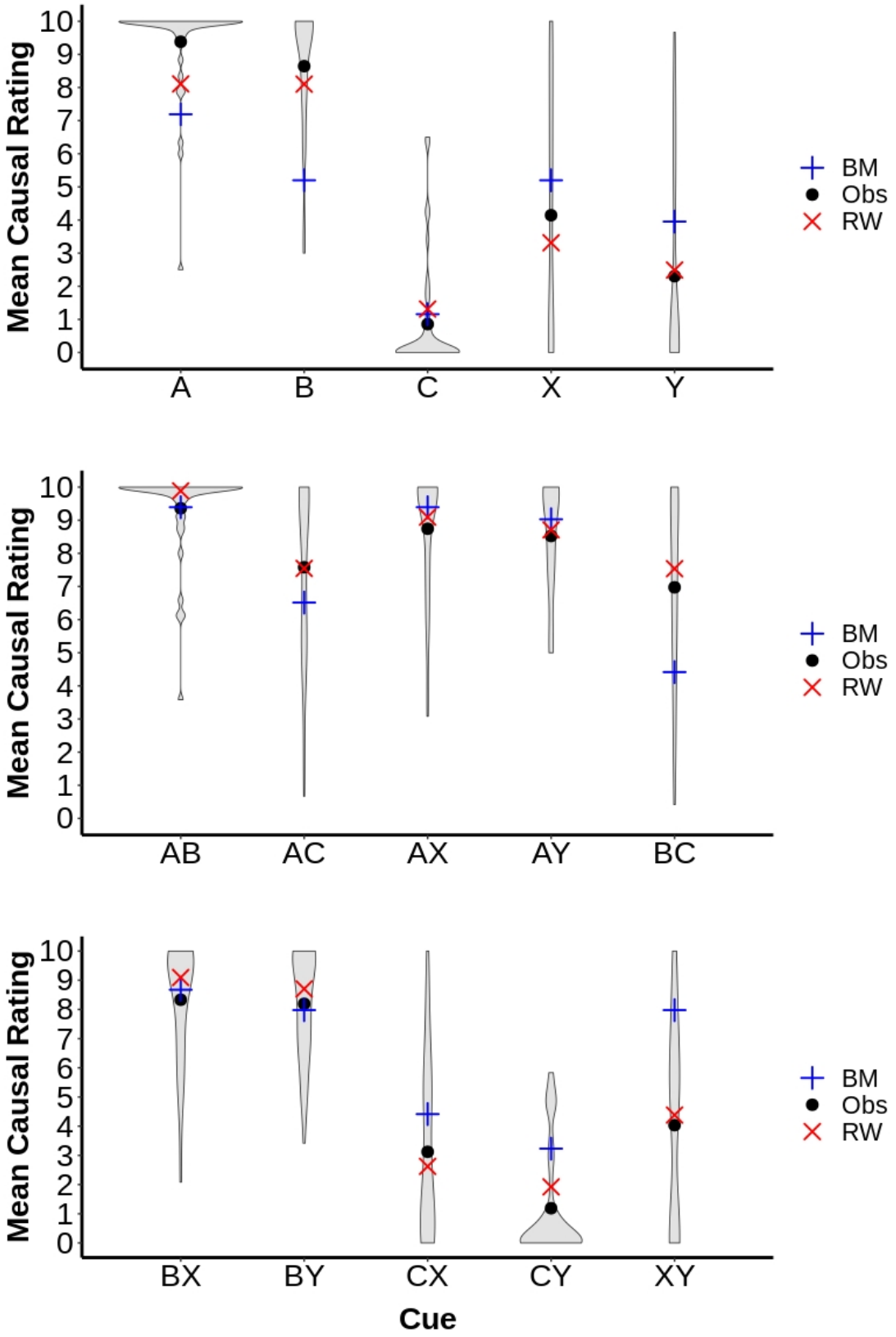

Figure 23. Predicted versus observed Test stage ratings for modified Rescorla-Wagner (RW) model, and Bush & Mosteller (BM) model, against observed data (Obs), following A+ AX+ BY+ CY- training. The violin plot represents the distribution of the observed data.

As with the blocking simulation, the best fitting initial associative strength was an intermediate value. It is notable that the value was slightly higher than for the blocking data. This could be because the outcome base rate during training (i.e. the proportion of trial types resulting in stomach ache versus no stomach ache) was higher. There is evidence from Jones et al. (2019) that participants' causal ratings of cues they are uncertain about are sensitive to the outcome base rate. If the value of the starting associative strengths is an adequate way of representing uncertainty about the causal status of novel cues, then the best fitting starting associative strength should change in line with the outcome base rate. It is possible to test this idea by model fitting on a dataset in which the outcome base rate has been experimentally manipulated. This was the basis of the final model fitting procedure.

4.4: Model fitting 3: redundancy effect base rate manipulation

A suitable dataset was already available for the final model fitting procedure. Jones et al. (2019) reported a redundancy effect experiment, in which the outcome base rate was varied between two different groups of human participants. Full experimental details are available in their paper and a brief summary is included below. The test stage likelihood ratings assigned to blocked cues were shown to vary in line with the outcome base rate, on the basis of participants being uncertain about the causal status of these cues. In both groups, the redundancy effect was observed, because blocked cue (X) was assigned higher ratings than the uncorrelated cue (Y). However, the rating for X was higher in the high base rate group. Consequently, the redundancy effect was larger when the base rate was higher. The manipulation was achieved by adding additional cues, so that either 25% or 75% of training trials resulted in stomach ache. To test the prediction that starting associative strength is sensitive to experimental base rate, this parameter was allowed to vary by condition in the model fitting procedure. None of the other parameters were allowed to vary by condition. If correct, the modified Rescorla-Wagner model should provide a good fit to all the test cues for both groups, with a higher best fitting initial associative strength in the 75% group than in the 25% group. As with the previous simulations, the training and test cues used for the model fitting were identical to the training and test cues experienced by the human participants. The fitting code for this simulation is available at https://osf.io/fh3gc/.

4.4.2: Base rate experimental details

The full details for this experiment are available in Jones et al. (2019). The trial-level raw data for this experiment is available at https://osf.io/fh3gc/. The design of the experiment is shown in Table 10. In this design, there are additional cues that are unrelated to the basic redundancy effect trial types (A+ AX+ BY+ CY-).

Table 10. The design of the Jones et al. (2019) redundancy effect base rate experiment

Stage 1 (75%)		Stage 1 (25%)		Test
	D+		D-	
	E+		E-	A
A+	F-	A+	F-	B
AX+	GH+	AX+	GH-	C
BY+	IJ+	BY+	IJ-	X
CY-	KL+	CY-	KL-	Y
	MN+		MN-	
	OP-		OP-	

Jones et al. (2019) predicted that that if participants are uncertain about the causal status of a cue, then they should assign their causal rating based on the overall frequency of the outcome. For example, if most training trials result in stomach ache, then participants should assign higher causal ratings to such cues, as it is more likely that the outcome will occur than not. They predicted that this between group difference would be observed for the blocked cue X but not the uncorrelated cue Y, since participants should be uncertain about the former but not the latter. The results showed a clear difference between the two groups, with a larger redundancy effect in the 75% group

compared to the 25% group. As predicted, the blocked cue ratings appeared to be labile, whilst the uncorrelated cue ratings were not.

4.4.3: Base rate model fitting details

The code for producing the model fitting is available at https://osf.io/fh3gc/. The model fitting process conducted on the redundancy effect base rate data used the same methodology as with the Experiment 8 blocking data and the Experiment 9 redundancy effect data. Unlike the previous datasets, the base rate data was taken from a between-groups experiment. The best fitting parameters again had to be the same across both groups, apart from the initial associative strength, since there was a theoretical reason for expecting this to vary in line with the different base rates in each experimental group.

Results from modified model implementation

The best fitting parameters, error and adequacy of fit are reported in Table 11. As predicted the best fitting initial associative strength was higher in the high base rate group than in the low base rate group.

Table 11. Output of modified Rescorla-Wagner model fitting on redundancy effect base rate data

LR	**Init Assoc (High)**	**Init Assoc (Low)**	**Beta**	**Theta**	**SSE**	**Mean Error**	**R^2 (Low)**	**R^2 (High)**
0.030	0.513	0.385	0.356	8.204	0.026	0.032	0.976	0.983

The modified Rescorla-Wagner model produced low error and an equivalently good fit in both the 25% and 75% base rate groups. As predicted, the best fitting initial associative strength was higher in the high base rate group than in the low base rate group. The initial associative strength parameter thus appears to provide one reasonable way of representing participants' uncertainty about the causal status of novel cues. Of course, real participants, unlike these simulations, need to experience at least a few trials in order to become sensitive to the outcome base rate. Thus, using initial starting weights to model the effects of outcome base rate is necessarily a simplification of the mental operations involved. Figure 24 shows the predicted versus observed Test stage ratings for the 25% and 75% base rate groups. The modified Rescorla-Wagner model was able to capture the redundancy effect in both conditions. The model was also able to capture the labile nature of the blocked cue X, although the effect of the base rate on X is slightly underestimated.

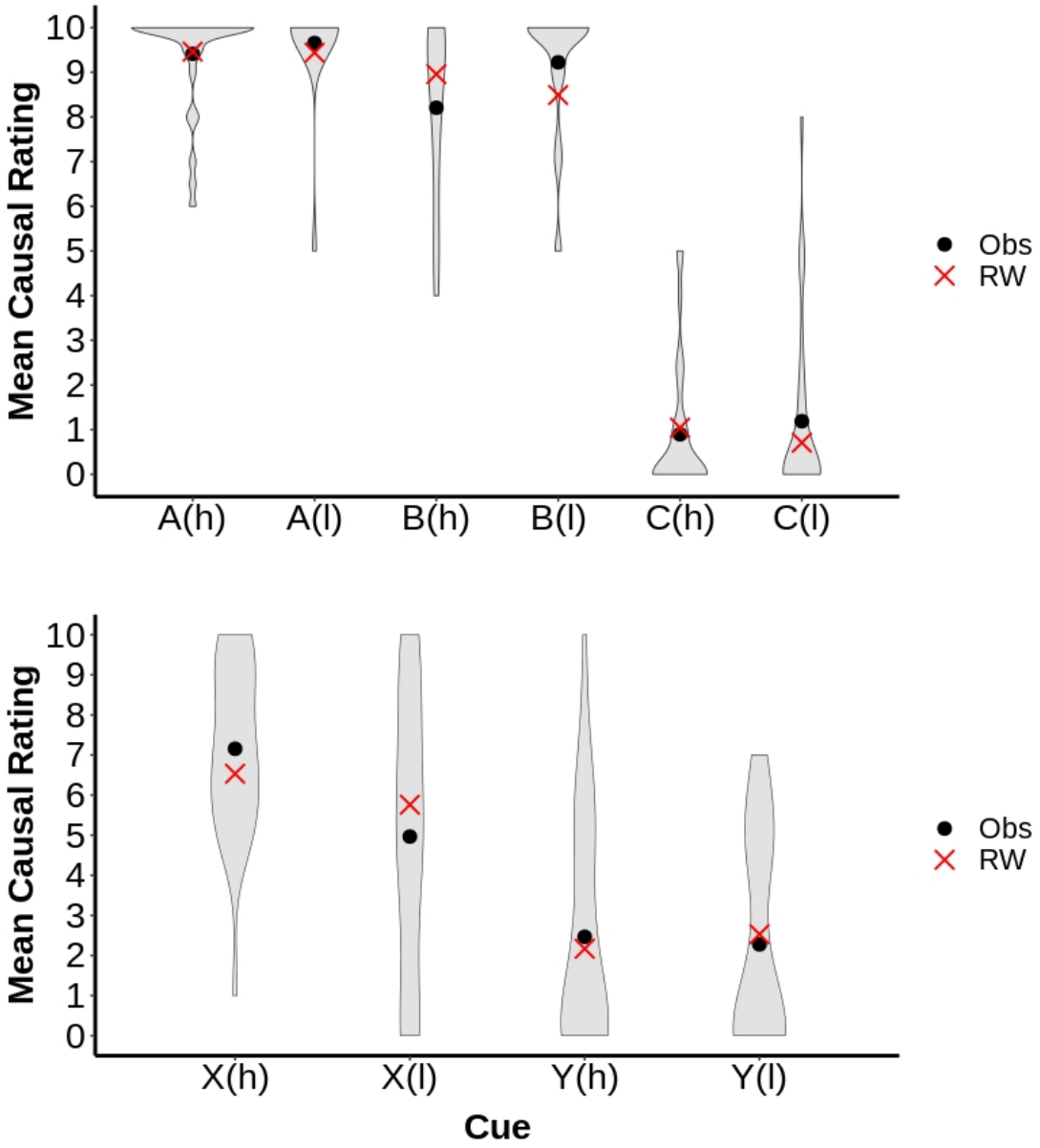

Figure 24. Predicted versus observed Test stage ratings for modified Rescorla-Wagner (RW) model, against observed data (Obs), fitting to the high (h) and low (l) base rate groups, following A+ AX+ BY+ CY- training (intermixed with additional cues used to manipulate the base rate). The violin plot represents the distribution of the observed data.

4.5: General Discussion

The assumption that associative strengths start at zero, in humans, is not necessarily correct. In fact, this assumption may be wrong in many common experimental scenarios. Allowing the initial associative strength to be an intermediate value provides a simple but effective way of formally representing participants' uncertainty about the causal status of novel cues. Contrary to intuition, the unmodified Rescorla and Wagner (1972) model provides no better an account of a standard forward cue-competition blocking experiment than the Bush and Mosteller (1951) model. However, if the Rescorla-Wagner model is modified, such that the initial associative strength of cues can be an intermediate value, then it does provide a better account than the equivalently modified Bush and Mosteller model. As stated in the chapter introduction (4.1), fitting models to whole experiments is important, because even though models might predict extracted parts of the experimental data, they should really be able to account for the full experimental data. The modified Rescorla-Wagner model is also able to adequately account for the redundancy effect (e.g. Uengoer et al, 2013), which the unmodified model cannot. In addition to capturing the effect itself, the modified model adequately captured the full observed test data using a complex set of cues. Furthermore, the initial associative strength of the simulated cues was shown to vary with the outcome base rate. The base rate finding supports the suggestion that intermediate initial associative strengths can be used to represent uncertainty about the causal status of novel cues. This is consistent with the experimental findings of Jones et al. (2019), in that the likelihood ratings assigned to cues with a uncertain causal status are influenced by the outcome base rate.

It is worth noting that Vogel and Wagner (2016) suggested an alternative modification of the Rescorla-Wagner model that also accommodates the redundancy effect. In their modification, common elements are added to the stimulus representation. While this modification has been shown to accommodate the basic redundancy effect, there is evidence that it cannot adequately account for the effect of varying the outcome base rate on the redundancy effect (Jones et al., 2019). Nevertheless, further investigation of both modifications, across a range of phenomena, would be a fruitful direction for future research. The importance of making broad relative adequacy comparisons of models has been previously emphasised within the literature (Wills & Pothos, 2012; Wills, O'Connell, Edmunds, & Inkster, 2017). Furthermore, it would be useful to know whether the initial associative strength modification allows other learning models to explain more phenomena in human learning. If this were the case, this modification might need to be considered when developing future models of associative learning applicable to humans.

Instead of model fitting to a redundancy effect dataset only incorporating five single cues at test, the fitting on the Experiment 9 data used an expanded set of test cues. Given the high adequacy of fit (R^2 value 0.96) observed for the modified Rescorla and Wagner (1972) model, a further fitting procedure conducted on a dataset incorporating only five single test cues could not produce a fit any worse than this. Whilst there is some scope for the Bush and Mosteller (1951) model to produce a better fit with less test cues, the best this could result in is both models producing a comparably good fit as each other. In this scenario, the modified Rescorla-Wagner model would still provide the best account across the data sets considered in this chapter. It was therefore concluded that it would not be worthwhile to conduct an additional model fitting

process on a redundancy effect dataset that only incorporates five single cues at test. Instead, it could be useful for future predictive learning experiments, which are going to be used for model fitting, to incorporate compounds at test, since this provides a more diagnostic method of testing the relative adequacy of models.

The simulations reported in this chapter demonstrate the importance of formal modelling, in that they show the capabilities of even very simple models can be hard to informally predict (as demonstrated by the blocking simulations with the unmodified models). Of course, informal predictions are sometimes proven correct, as with the redundancy effect simulation using the unmodified Rescorla-Wagner model. Nevertheless, these results clearly demonstrate the value of thoroughly investigating the parameter space of formal models, and highlight the potential risks of relying solely on informal intuition. Formal simulations are becoming more common in the study of human predictive learning, and have been well established in related fields, such as category learning, for decades. One possible barrier to conducting model simulations is the perceived high entry cost, and the apparent lack of a common open framework, in which models, phenomena and simulations can be easily assessed and compared. However, options are available, such as ALTSim (Thorwart, Schultheis, König, & Lachnit, 2009). ALTSim does not allow for parameter space optimisation, but it does allow initial associative strengths to be set to values other than zero. The model implementations reported in the current chapter used a free and open source package called *catlearn* (Wills, Dome, Edmunds, Honke, Inkster, Schlegelmilch, & Spicer, 2019), which is available to download in the open-source R environment (R Core Team, 2018). Catlearn includes a number of model implementations, including Rescorla and Wagner (1971), Bush and Mosteller (1951), EXIT (Kruschke, 2001), and COVIS

(Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Catlearn is an extensible framework and it is easy to add more models (by request or through distributed collaboration), and to contribute to the project. You can find out more information here: https://ajwills72.github.io/catlearn/

The initial strength modification explored in this chapter is just one feature that might be useful to incorporate into an associative learning model based on theory protection. Indeed, the intermediate starting associative strength of novel cues in this chapter is consistent with the intermediate ratings given to the novel cue at test in Experiment 4 (Chapter 3). Additionally, the concept of uncertainty about novel cues is analogous with the idea of having low confidence (i.e. a weak theory) about cues with an ambiguous causal status. Of course, not all cues with an ambiguous causal status will be assigned intermediate causal ratings by participants. For example, Experiments 6 and 7 showed participants assigning low ratings to cues with an ambiguous status, because their uncertainty encompassed a restricted range of possibilities (i.e. they had enough information to learn that non-reinforced cues cannot be a cause of the outcome). Furthermore, the theory protection account appears to be about more than causal confidence. For example, Chapter 2 briefly discussed how theory protection may operate when there is no difference in confidence between cues. Also, Chapter 3 introduced the idea of matched (i.e. consistent) causal statuses and outcomes in theory protection. The final chapter discusses these issues in greater detail, and sets out a follow-up programme of research for developing theory protection into a formal model.

Chapter 5

5.1 Implications of the research findings

At present, there is enough evidence to warrant further investigation of the theory protection account. The results of the three experiments reported in Chapter 2 are incompatible with either individual (Bush & Mosteller, 1951) or overall (Rescorla & Wagner, 1972) prediction error, also ruling out a hybrid of these (Rescorla, 2001) as an explanation. They are also not compatible with the attentional model proposed by Pearce and Hall (1980). However, they could be accounted for by Mackintosh's (1975) attentional model. The results of the four experiments reported in Chapter 3 are also inconsistent with a prediction error account. The first two experiments are explainable according to Pearce and Hall's, and Mackintosh's attentional models, but the subsequent two experiments were not consistent with either of these models. In short, theory protection is the only account considered so far that is able to account for all of Experiments 1-7. Furthermore, Experiment 7 demonstrated a difference in confidence about the causal status of cues that is consistent with theory protection. The theory protection account is also consistent with learning during Stage 2 of Experiment 8 reported in Chapter 4. Recall that cue B from the feature positive control part of the design was trained as a non-reinforced cue (B-) in Stage 1, before being placed in a causal compound (BY+) in Stage 2. The higher likelihood rating assigned to Y than B at test is consistent with participants protecting their theory about B not being causal. Instead, participants attributed the outcome on BY+ trials to the novel cue Y. The design of the other experiments used for model fitting, in Chapter 4, did not allow for any

assessment of theory protection, because there was only one training stage in each experiment.

Taken as a whole, the results of Experiments 1-8 provide evidence that theory protection plays a prominent role in selectivity in human learning. That is, participants maintain previously learned causal associations, instead attributing unexpected outcomes to cues with a comparatively ambiguous status. Furthermore, the possible causal status range of cues (on the basis of prior training), and whether subsequently encountered outcomes match that causal status range, also appears to be central to theory protection. In the current experiments, when two cues were trained in compound, participants always learned more about the cue with the causal status range that matched the current trial feedback (i.e. outcome). Conversely, participants protected their theory about the cue with the causal status that did not match the current trial feedback. For example, at the end of Experiment 6 Stage 1, the status of cue B following B- training could either have been neutral or preventative (but not causal). Therefore when B was paired with causal cue A (following A+ training) during Stage 2, in a non-reinforced compound AB-, the status of cue B matched the current trial feedback. Note that an inhibitory cue would be able to prevent A from causing the outcome. However, the status of A did not match the Stage 2 feedback. To provide another example, at the end of Experiment 1 Stage 1, the status of the blocked cue X could have either been causal or non-causal, while the uncorrelated cue Y should not have been considered a cause. Therefore the possible status range of cue X matched the outcome during the XY+ trials, while the status of Y did not.

This principle applies to all of Experiments 1-8, because the cue that was learned about the most in Stage 2 encompassed a range of possible causal statuses following Stage 1 training. Part of this range always matched with the outcome that was encountered during Stage 2. This range was a consequence of the experimental manipulations causing participants to lack confidence about the status of those cues. The cue that was learned about the least in Stage 2 of experiments 1-8 always had a causal status that did not match with the outcome encountered during Stage 2. However, as already outlined, the theory protection account also makes predictions for designs in which there should be no lack of confidence about the causal status of cues trained in compound. Therefore, while the focus of these experiments has been on causal confidence, matching between causal statuses and outcomes appears to be equally (if not more) important to theory protection in associative learning. It is possible that causal ambiguity, and the resulting lack of confidence it causes, drives learning because it facilitates matching. Some experiments intended to test these ideas are proposed in the next section of this chapter.

Contrary to what one might assume from a prediction error account, the evidence presented in this thesis shows that the cue with the greatest prediction error is often the one least learned about. Instead, it appears that humans resist changing strongly held beliefs. People seem to attribute surprising outcomes to cues when they either lack a strong theory about that cue, or when the theory they do hold matches the surprising outcome. In other words, when protecting their theories about cues, humans should look for alternative causes that might provide a better explanation of the outcome, such as cues with a lower prediction error. Despite the apparent inconsistency between a prediction error account and the results of the present experiments, the idea that prediction error influences learning should not be dismissed in its entirety. Broadly

speaking, learning seems to be dependent on there being a discrepancy between predicted outcomes and actual outcomes. In the theory protection account, this 'surprise' may still dictate whether learning takes place, and how much learning there will be. Meanwhile, the theories that people develop (and which they subsequently protect) about the causal status of cues may dictate which cues are learned about, in order to predict future events more accurately. All of Experiments 1-8 introduce a surprising outcome (or a surprising outcome omission). For example, in the Chapter 3 experiments, a previously causal cue and an ambiguous cue were paired in a non-causal compound in Stage 2. In each of these experiments, the discrepancy between the stomach ache predicted by A, following Stage 1, and the absence of stomach ache following the Stage 2 AB- compound presumably led to learning about B. It is likely that, in the absence of surprise, no learning would have taken place. This is an important point to consider for the purpose of developing a formal model of theory protection in associative learning.

Chapter 4 investigated uncertainty; specifically, how uncertainty about novel cues could be represented within learning models. The results from a set of model fitting simulations challenge the assumption that the associative strength of cues should start at zero in human learning. This assumption appears to be incorrect in at least some human predictive learning tasks. Future simulations of human learning could benefit from setting the initial associative strength of cues to an intermediate value, to represent participants' lack of confidence about the causal status of cues not yet encountered. Making this modification to a simple learning model (i.e. Rescorla and Wagner, 1972) allows it to accommodate more phenomena than the unmodified model; a simple blocking experiment, the redundancy effect, and the effect of outcome base rate on the

redundancy effect. Further investigation of the starting associative strength modification, across a wide range of models and phenomena, would be a useful next step in representing uncertainty in the modelling of human learning. It may be appropriate for future models of human learning to incorporate this parameter. This could include any potential future model of theory protection in associative learning.

A formal model of theory protection would need to represent the range of possible causal statuses that a cue might encompass. For example, a blocked cue could be causal or neutral, but not inhibitory. A non-reinforced cue could be either neutral or inhibitory, but not causal. A cue where the outcome is concealed could have any status, assuming this is permitted by the experimental scenario. Meanwhile, a reinforced cue that is trained in isolation could be causal, but could not have any other status. A model of theory protection would also need a process to calculate which cues are learned about across a wide range of experimental designs. This would require the mathematical representation of matching between the causal status (or causal status range) of cues and subsequently experienced outcomes. Finally, the amount of learning would also need to be calculated. This last point is perhaps the most straightforward, as this would presumably be calculated from a form of overall prediction error. As outlined above, some kind of discrepancy between predicted and experienced outcomes should still be needed for learning to take place. The causal status range of cues could be achieved by having an upper and lower bound to associative strengths, representing the limits of what causal status a cue could have. For example, a cue where the outcome is concealed would have a range encompassing all causal statuses; an upper bound of one, and a lower bound of minus one. A non-reinforced cue would have an upper bound of zero and a lower bound of minus one. Meanwhile, a cue with an unambiguously known

status would have an upper and lower bound at the same associative strength value. Importantly, because ambiguous cues are subsequently trained in compounds, in which their causal status can be ascertained, the upper and lower bound would reduce on a trial-by-trial basis. The representation of matching is a little less clear at this stage, but it might involve subtracting the associative strength value of an experienced outcome from all possible values within the associative strength range of a cue. For example, if this calculation produces a value of zero during such a process, then this could be a condition under which that cue is learned about. Of course, these suggestions are highly speculative at this stage.

The Bayesian approach to associative learning (e.g. Kruschke, 2008) outlined in the introduction could also provide a way of implementing the types of processes suggested above. The belief distribution used to represent the status of cues may provide a logical way of capturing the causal status ranges described above. Additionally, the weighting of competing hypotheses (about the status of cues) could provide a way for beliefs about ambiguous cues to update faster than beliefs about cues with a causal status that is more confidently known. In other words, existing cue-outcome associations could be protected if participants have a high degree of confidence, in circumstances where cues are trained together in compounds. As mentioned in Chapter 1, a model using this approach would need learning to be cue-governed, rather than outcome-governed. Before attempting to construct a testable formal model, using one of the approaches suggested, it is necessary to conduct further experiments to investigate some of the additional processes that may underpin theory protection.

Before outlining follow-up experimental designs, there are some further points that briefly require discussion. For example, the findings in Experiments 6 and 7 were the opposite of those found in rats and pigeons by Rescorla (2001). Moreover, the findings of Experiments 1-8 were the opposite of the results predicted by the Rescorla (2001) account. This suggests that further comparative studies between humans and non-human animals would be valuable. The issue of whether a comparable process to theory protection occurs in non-human animals needs addressing. At present, it is tempting to conclude that there is no such process in rats and pigeons, on the basis of Rescorla's results. However, many of the experimental designs reported in this thesis were different to those conducted by Rescorla. It is therefore possible that close replications of the designs reported here, using non-human animals, would produce equivalent results to humans. Furthermore, it is possible that other species of animals do in fact learn in a way that is more human-like, with respect to theory protection. Of course, such inter-species differences may not be as straightforward as they initially seem. For example, there is already some evidence (i.e. Experiments 6 and 7 from this thesis compared with Haselgrove and Evans, 2010) that humans learn more like rats and pigeons if the scenario is appropriately manipulated.

Attention has not been totally ruled out in explaining the current set of results. The potential role of attention could be further tested by examining overt attention to cues during training. It should be possible to detect whether there is a difference in the attention paid to the cues at the start of Stage 2 (in Experiments 1-8) by tracking eye movements. Furthermore, if there is a rapid shift in attention from one cue to another during Stage 2, it should also be possible to detect this from eye-tracking. It may be that both theory protection and attention have a role to play in human associative learning,

rather than these processes being mutually exclusive. For example, it could be that matching (between cue status and outcome) and variations in confidence (about the causal status of cues) result in differences in the attention paid to such cues. Furthermore, even if attention correlates with learning, it may not govern it. Both learning and attention could be driven by the theories people hold about cues, on the basis of the information available to them while they learn.

In summary, the findings in humans reported here have three potential implications: that the learning mechanisms responsible for rat behaviour might be different to those found in humans; that the modelling of human behaviour may require the development of a new theory; that the operation of prediction error (as currently conceptualised) might need modifying. The remainder of this chapter proposes a further programme of research investigating the cognitive processes and mechanisms required to explain theory protection in human associative learning. The findings from these proposed experiments will aid in the development of a new model of learning.

5.2: Towards a formal model of theory protection

Further experimental investigation of theory protection in human learning is required, so that a formal model can be developed. At present, an informal account has been proposed, in which humans (unlike certain non-human animals) resist updating existing cue-outcome associations as much as possible, when faced with surprising outcomes. In order to turn this informal account into a formal theory, it is necessary to further unpack the cognitive processes underpinning theory protection in human associative learning. There are several avenues for future research that will achieve this aim.

The experiments presented in Chapter 3 show participants learning more about an ambiguous cue than an unambiguous cue, when both are trained in a compound that is not causal. It appears that the ambiguous cue is learned about more because the range of causal statuses it could encompass at the end of Stage 1 matches the outcome experienced during Stage 2. In each experiment, it was possible that the ambiguous cue (B-) was an inhibitor. The combination of the cues and the outcome during the compound stage (AB-) was sufficient for participants to learn that B was an inhibitor with a strong degree of confidence. The next step is to further test how this generalises. For example, will this effect work in the opposite direction, so that participants learn more about a cue that might be causal and protect their theory about a known inhibitor? The reason for suggesting an inhibitory cue is that an unambiguous inhibitor is the causal opposite of an unambiguous excitor. This could be tested in a series of experiments that mirrors the three basic designs used in Chapter 3 (note that Experiments 6 and 7 are effectively the same design). The first experiment would therefore contain an inhibitor trained in a causal compound with a novel cue, mirroring the causal cue trained in a non-causal compound with a novel cue in Experiment 4. The

subsequent experiments would follow the same logic, except that the second experiment would use a concealed-outcome cue and the third experiment would use a blocked cue. In the case of a blocked cue, this could either be causal or neutral, but not inhibitory, which is the opposite of a non-reinforced cue (which could either be neutral or inhibitory, but not causal).

Table 12. Design of three experiments to test the generalisability of theory protection

Experiment	Stage 1	Stage 2	Test
A	A+ C+ AM- CN-	MB+	NB MD A B C D N M
B	A+ C+ AM- CN- B? D?	MB+	NB MD A B C D N M
C	A+ C+ AM- CN- E+ EB+ F+ FD+	MB+	NB MD A B C D E F N M

Key:
Letters = different cues
+ = outcome
- = no outcome
? = outcome concealed from participants

In the Table 12 experimental designs, participants should learn more about the ambiguous cue B according to theory protection, but should learn more about the inhibitor M according to a prediction error account (e.g. Rescorla, 2001). This is because Cue M would be most discrepant with the Stage 2 outcome. According to Mackintosh (1975), cue M would also have the higher associability at the start of Stage 2 in each of the above designs (assuming that novelty does not result in higher associability in Experiment A). This is because M would be a good predictor of the absence of the outcome. The theory protection account predicts more learning about B, on the basis of participants protecting their belief that M is preventative of the outcome

in each design. Logically, participants should learn that B is a strong cause of the outcome (i.e. B should become superconditioned), in order to overcome an inhibitor. The aforementioned results of Le Pelley and McLaren (2001) showed more learning about a causal cue than an inhibitor, in a comparable design, although their experiment did not incorporate any causal ambiguity.

A current gap in the understanding of theory protection concerns extinction and reacquisition. In typical prediction error models, learning is represented as starting off fast for cues that are trained as causal, assuming an initial associative strength of zero (although this associative strength assumption may be wrong in humans as indicated by the findings of Chapter 4). As the prediction error reduces, learning slows until the asymptote of learning is reached. If those same cues are subsequently extinguished, learning should again start off fast and then slow down as extinction reaches asymptote. If those cues are then re-trained as causal, learning should once again start off fast and become slower. However, if participants engage in theory protection, then they should initially resist updating the associative strength of cues with a known status. This initial resistance to updating would obviously not apply to cues being trained for the first time, but it would apply to any subsequent extinction and re-acquisition. According to theory protection, learning should start off slow, become faster as the current theory about those cues is released, and then slow down again as learning reaches asymptote. Recall the Chapter 1 example of watching a disappointing new season of a previously brilliant TV show, for an everyday example of how such a process could manifest. Your prediction about the quality of subsequent episodes would change rapidly from the outset according to a prediction error account, but you would resist such initial rapid change in your predictions, according to the theory protection account. The design in

Table 13 outlines a possible way of detecting such a learning pattern. As before, the use of compounds is intended to overcome the potential non-linear mapping between associative strength and responding. Four cues are trained as causing an outcome (e.g. happiness) in Stage 1. Two of those cues (B and D) are then trained as causing a mutually exclusive outcome on the same scale (e.g. sadness) in Stage 2. Cue B is then trained, along with cue A, in a compound that causes this second outcome, in Stage 3. If learning starts off slow and then speeds up, this should result in more learning about B than A in Stage 3 (assuming the rate of learning is not so fast that learning about B nears completion in Stage 2). Consequently, BD should be rated sadder than AC at test according to this conceptualisation of theory protection. However, according to a prediction error account, AC should be rated sadder.

Table 13. Extinction/reacquisition experimental design

Experiment	Stage 1	Stage 2	Stage 3	Test
D	A-O1 B-O1 C-O1 D-O1	B-O2 D-O2	AB-O2	AC BD A B C D
	Key: Letters = different cues O1 = outcome 1 O2 = outcome 2			

As already noted, the experiments presented in this thesis predominantly focused on differences in confidence about the status of cues, but matching between the status of cues and subsequent outcomes needs further investigation. For example, more needs to be understood about situations in which there is no difference in participants' confidence about the status of cues trained in compound. Participants might be equally confident about two cues, but the theory protection account should still make

predictions about which cue will be learned about the most, on the basis of existing causal knowledge. A useful starting point might be to replicate the Le Pelley and McLaren (2001) experiment described near the end of Chapter 2. This experiment has not yet been replicated and there were only 20 participants. As a useful follow up to this, two additional experiments using novel designs are outlined in Table 14. As before, the outcomes (O1 and O2) would need to be mutually exclusive.

Table 14. Design of two experiments investigating matching

Experiment	Stage 1	Stage 2	Test
E	A-O1 B-O2 C-O1 D-O2	AB-O1	AC BD A B C D
F	XA-O1 YB-O2 XC-O1 YD-O2 X-O1 Y-O2	AB-O1	AC BD A B C D X Y

Key:
Letters = different cues
O1 = outcome 1
O2 = outcome 2

In Experiment E, the design tests for matching influencing selectivity in learning. Confidence about the causal status of the cues should not differ at the end of Stage 1, so that participants are equally confident about the status of A and B. During Stage 2, the previously learned status of A matches the outcome that is trained with the AB compound, but the previously learned status of B does not match. Participants should therefore protect their theory about B, instead learning more about A and assigning a rating that reflects this to AC at test. According to a prediction error account, the opposite result should be seen, as the associative strength of B would be most discrepant with the outcome during Stage 2. One possible flaw with this design is that it may not be possible for the associative strength of A to increase any more (during Stage 2) if

learning is at asymptote. This could be overcome by presenting a stronger outcome during Stage 2. Experiment F follows the same logic, except that participants should have an equal lack of confidence about the causal status of both A and B following Stage 1. Cues A and B are both blocked cues and are therefore causally ambiguous from the perspective of participants (Jones et al., 2019). As before, cue A has the better matched outcome during Stage 2, so according to the theory protection account there should be more learning about A, assuming matching is important to learning.

The two experimental designs in Table 14 raise a follow-up question about which of matching or confidence is the most important aspect of theory protection. In Experiments 1-8, the cue with the matched outcome was also the one about which participants were least confident, which makes it impossible to answer that question from these datasets. The two designs in Table 15 effectively pit confidence and matching against each other, in order to address this issue. In Experiment G, causal ambiguity is maximal for cues C and F in both groups, because the outcome is concealed during Stage 1. In Group 1, A is better matched than C in Stage 2, but participants should be less confident about C than A after Stage 1. Therefore, according to matching, there should be more learning about A during Stage 2, but according to confidence, there should be more learning about C. In Group 2, C is better matched than A during Stage 2 (and also more ambiguous), but A would have the greater prediction error. Therefore, according to matching there should be more learning about C. In short, if confidence is dominant, then C should be learned about most during Stage 2, in both groups. However, if matching is dominant, then A should be learned about most in Group 1, while C should be learned about most in Group 2.

Table 15. Testing matching against confidence in theory protection

Experiment	Stage 1	Stage 2 (Group 1)	Stage 2 (Group 2)	Test
G	A-O1 B-O2 C-? D-O1 E-O2 F-?	AC-O1	AC-O2	AF DC
H	A+ B- C+ D-	AB+	AB-	AC BD
	Key: Letters = different cues O1 = outcome 1 O2 = outcome 2 + = outcome - = no outcome ? = outcome concealed from participants			

Experiment H follows a similar logic, except that causal ambiguity is across a more restricted range for cues B and D, because participants can at least learn that these cues are not causal during Stage 1. In Group 1, A has the matched outcome during Stage 2, while B has the most ambiguous status at the end of Stage 1. According to matching, there should be more learning about A, but according to confidence (and prediction error) there should be more learning about B. In Group 2, B has both the matched outcome and the most ambiguous status, while A would have the greater prediction error. According to theory protection (both matching and confidence) there should be more learning about B during Stage 2 (which of course has already been observed in Experiments 6 and 7 in this thesis).

The Mitchell et al. (2008) experiment described in Chapter 2 is of some relevance to the relationship between confidence and matching. Recall that a previously causal cue (A+) and a novel cue (B) were subsequently trained in a causal compound (AB+) resulting in evidence of more learning about the novel cue. Taken a face value, this result suggests

that confidence is dominant over matching, since A was the better matched cue during the compound training, while B was causally ambiguous. However, the range of outcomes that B could be, as a novel cue, includes the possibility of it being causal. Therefore, it could be argued that both A and B match. Group 1 in Experiment H overcomes this issue, because B would not match the Stage 2 outcome, as a consequence of it being non-reinforced in Stage 1. As for the novel cue being learned about most in the Mitchell et al. design, it is possible that participants inferred B was not inhibitory (even if it could still have been either neutral or causal) following feedback on Stage 2, since it would have prevented the outcome during this stage if so. However, the scenario used foods as the cues, so it seems unlikely that participants would have considered that B might be inhibitory in the first place. Furthermore, the presence of a prediction error to drive learning during the compound stage is unclear. Cue A already predicted the outcome during the first training stage, so the occurrence of the outcome during the compound stage should not be surprising, although participants might over-predict the outcome, setting up a negative prediction error. Whilst the Mitchell et al. result seems broadly consistent with theory protection, the exact mechanism of operation is hard to identify. The other experiments proposed here should hopefully allow for some clarity. It may be that neither matching nor confidence are dominant, and that other factors might guide which of these processes drives theory protection across different experimental designs.

The Chapter 3 discussion (3.6) raised the point of what would happen in cases where it is not possible for participants to protect their existing theories about cues. In other words, what happens in cases where something must have changed? One possibility is that learning would resemble individual prediction error, although this learning may be

slow at first if participants resist updating their beliefs. Again, think back to the example of a disappointing new season of a previously brilliant TV show, for a simple real-life example of when something must have changed. Another way of framing this, in the context of compound learning experiments, is to consider what would happen when the status of neither cue in a compound matches an experienced outcome. The experimental designs in Table 16 will allow this to be tested.

Table 16. Design of two experiments investigating learning when theory protection is not possible.

Experiment	Stage 1	Stage 2	Test
I	A+ AX- Y- B+ BW- Z-	XY+	XZ WY
J	A+ AX- Y- AY+ B+ BW- Z- BZ+	XY+	XZ WY
	Key: Letters = different cues + = outcome - = no outcome		

In Experiment I, neither X nor Y match the outcome during Stage 2, since X is trained as an inhibitor during Stage 1, while Y is non-reinforced. Y is more causally ambiguous, but participants should have learned that it is not a cause of the outcome during Stage 1, even if it might be either neutral or inhibitory. Therefore, this difference in confidence between cues X and Y should not result in more learning about Y, if one assumes that matching is integral to theory protection. Meanwhile X will have an unambiguous causal status following Stage 1. If participants are unable to protect their theory about either X or Y, then it is possible that there will be more learning about X than Y (resembling individual prediction error), as participants are forced to release their theory

about X being inhibitory. Of course, it is also possible that participants would learn about both cues equally (resembling overall prediction error). Experiment J follows the same logic, except that Y and Z are both unambiguously trained as neutral cues, thus removing any lack of confidence about their causal status as a variable that might govern learning. It should be noted in these designs, that Y is arguably the 'closer' matched cue in Stage 2, so if there was more learning about Y than X, then that is perhaps one possible interpretation of such a result. However, neither cue should be an adequate match, so theory protection (as defined in this thesis) should still not be possible.

Another possible design mentioned in the Chapter 3 discussion (3.6) is included in full, in Table 17. Similarly to Experiments I and J, this experiment is intended to test what will happen if participants are unable to protect existing theories. In Group 1, B is a blocked cue, while in Group 2, the design from Experiment 6 is replicated, so that B is non-reinforced. Theory protection should not be possible in Group 1, because a blocked cue cannot be inhibitory, otherwise the outcome would not be observed on the EB+ trials. Therefore, unlike Group 2, there should not be greater learning about B than A during Stage 2 in Group 1.

Table 17. Design of experiment investigating learning when theory protection is not possible

Experiment	Stage 1 (Group 1)	Stage 1 (Group 2)	Stage 2	Test
K	A+ C+ E+ EB+ F+ FD+	A+ C+ B- D-	AB-	AD BC
	Key: Letters = different cues + = outcome - = no outcome			

The Chapter 3 discussion (3.6) also suggested a between-groups experiment, intended to unpack the differences between the theory protection account and the Pearce and Hall (1980) attentional model. The design is included in full, in Table 18. In Group 1, the design of Experiment 6 is replicated, but with an extra pre-training stage containing some filler cues. In Group 2, B- is explicitly pre-trained as inhibitory. The theory protection account predicts a lower rating for BC in Group 1, but no difference between the test compounds in Group 2, resulting in a between-group interaction. Pearce and Hall's model predicts no difference between the groups because all cues should decline in associability equally before Stage 2. This is because the outcome should be equally predictable on all trials.

Table 18. Design of experiment investigating theory protection and Pearce and Hall (1980) attentional model

Experiment	Pre-Train (Group1)	Pre-Train (Group2)	Stage 1	Stage 2	Test
L	P+ Q+ R- S-	P+ Q+ PB- PD-	A+ C+ B- D-	AB-	AD BC
	Key: Letters = different cues + = outcome - = no outcome				

Finally, the Chapter 3 discussion (3.6) suggested that theory protection could be ameliorated by manipulating the scenario between food-allergy and chemical-allergy. As stated, there is already some indirect evidence that theory protection is dependent on people's assumptions about the scenario (i.e. Haselgrove and Evans (2010) findings versus the results of Experiment 6 in Chapter 3). In short, the food-allergy scenario

appears to restrict the type of learning that can occur, because participants do not generally assume that foods can become inhibitory (also supported by Zaksaite & Jones, 2019). This idea relates back to matching. In the Haselgrove and Evans experiment, matching between the non-reinforced cue from Stage 1 and the absence of the outcome during Stage 2 should not have been possible, because the status of the non-reinforced cue would not have extended to it potentially being an inhibitor (i.e. participants should have learned that it was neutral with a high degree of confidence). Such assumptions about the scenario could be regarded as another form of theory that people protect, in spite of surprising outcomes. In order to protect such theories, participants would have to release competing theories about the causal status of specific cues. The role of the scenario would need to be investigated using between-groups experiments, in which one group is given a chemical scenario and another group is given a food scenario. The three designs from Chapter 3 (please see Table 2) would be suitable.

5.3: The basis of theory protection

Whilst the focus of this thesis has mainly been on the processes underpinning theory protection in human associative learning, there is also the question of what cognitive basis such processes might have. If theory protection is unique to humans (which it might not be), it may be that it relies on complex, abstract cognition, as opposed to a low-level, automatic associative mechanism. If this were the case, then it would make sense for such cognitive processes to be unique to humans, or at least unique to species capable of more complex forms of cognition, such as primates, dolphins, or pigs (e.g. Tomasello & Call, 1994; Herman, 2006; Marino & Colvin, 2015). In humans, the potential involvement of complex cognition in theory protection could be investigated using cognitive load manipulations. The basic idea is that if participants are completing a learning experiment in which theory protection appears to operate, then it should be possible to ameliorate theory protection by giving participants a simultaneous task that occupies their cognitive capacity. Such experiments would need to be between-groups, with one group incorporating the cognitive load manipulation, and a control in which no cognitive load is introduced.

The three experimental designs in Chapter 3 would be suitable for a cognitive load manipulation, as the effect seen across those experiments seems particularly robust. The designs can be found in Table 2. There are several ways in which cognitive load can be introduced to experimental procedures. For example, Wills, Graham, Koh, McLaren and Rolland (2011), and Seabrooke, Wills, Hogarth, and Mitchell (2019), introduced cognitive load by requiring participants to memorise numbers while completing learning experiments. Abelson, Erickson, Mayer, Crocker, Briggs, Lopez-Duran, and Liberzon

(2014) created cognitive load through stress, by requiring participants to prepare for giving a talk. Other ways of creating cognitive load include adding a time constraint to learning tasks, and asking participants to count backwards in intervals (e.g. Moghadam, Ashayeri, Salavati, Sarafzadeh, Taghipoor, Saeedi, & Salehi, 2011). It may also be possible to introduce cognitive load through anxiety. For example, the inhalation of air rich in carbon dioxide can be used to create a temporary state of anxiety in participants (e.g. Levitt, Brown, Orsillo, & Barlow, 2004).

5.4: The development of theory protection

A final avenue for potential future research is investigating the development of theory protection in learning, to find out at what stage humans begin protecting theories in this way. This links with using cognitive load to investigate whether complex cognition is involved, because children develop specific cognitive abilities at different stages of development, such as developing a theory of mind (Mitchell, 1997). This also links with investigating the role of the scenario, because children under a certain age may not have developed theories such as foods not being inhibitors of allergic reactions. Studying the developmental trajectory of theory protection is interesting for three reasons. Firstly, it could tell us about which areas of the brain (and which cognitive processes) are important. Secondly, there may be potential applied applications for this kind of research, such as for teaching and education. Thirdly, it would be interesting to know if humans start out ‘rat-like’, learning via low-level mechanisms, before gaining theory protection as part of a suite of more complex cognitive abilities. A selection of the experimental designs covered in this thesis could be adapted and given to children of varying age groups, to test for evidence of theory protection.

5.5: Concluding statement

In conclusion, the product of this thesis is an informal account of theory protection in human associative learning, based on a series of experiments that support this view. As discussed, the findings of these experiments are irreconcilable with the predictions of several major theories of associative learning. Additionally, this thesis proposes a simple mathematical way of representing human participants' lack of confidence about the causal status of novel cues within existing models of associative learning. Future research into theory protection should use a combination of theory-driven experimentation and formal computational model fitting to test these ideas further. However, before any formal modelling of theory protection can be conducted, it is first necessary to run some of the additional experiments covered in this chapter, so that a testable formal model of theory protection in human associative learning can be developed. In particular, the experiments looking for evidence of matching, the role of the scenario, and the generalisability of theory protection will serve this purpose. However, even in the absence of a formal model, the evidence for theory protection in learning is compelling, and could have implications for future associative learning research. In particular, the way in which human participants appear to protect associations from change should be of greater focus for future research in this field.

Addendum: equations for attentional models

According to Mackintosh's (1975) attentional model, more attention is paid to cues that are better predictors of outcomes, resulting in greater learning. Attention (represented as associability) operates alongside individual prediction error. The equation is as follows:

$$\Delta V_x = S\alpha_x(\lambda - V_A) \quad (4)$$

In Equation 4, S is the learning rate parameter and α is the associability. The associative strength is denoted by V, where ΔV_x is the change in associative strength for cue X, and V_x is the current associative strength of cue X. The asymptote of learning is represented by λ.

According to Pearce and Hall's (1980) attentional model, more attention is paid to cues that are followed by surprising outcomes, while cues that are followed by predicted outcomes decline in associability. The equation is as follows:

$$\Delta V_x = S\alpha_x\lambda \quad (5)$$

In Equation 5, S is the learning rate parameter, and the associability of cue X is represented by α_x. As before, ΔV_x is the change in associative strength for cue X, and λ is the asymptote of learning. The associability is calculated by subtracting the sum of the associative strengths on the previous trial from λ.

References

Abelson, J. L., Erickson, T. M., Mayer, S. E., Crocker, J., Briggs, H., Lopez-Duran, N. L., & Liberzon, I. (2014). Brief cognitive intervention can modulate neuroendocrine stress responses to the Trier Social Stress Test: Buffering effects of a compassionate goal orientation. *Psychoneuroendocrinology*, *44*, 60-70.

Ashby, F. G., Alfonso-Reese, L. A., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Baguley, T., & Kaye, D. (2010). Book review: Understanding psychology as a science: An introduction to scientific and statistical inference. *British Journal of Mathematical and Statistical Psychology, 63*, 695–698.

Balooch, S. B., & Neumann, D. L. (2011). Effects of multiple contexts and context similarity on the renewal of extinguished conditioned behaviour in an ABA design with humans. *Learning and Motivation*, *42*(1), 53-63.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.

Beckers, T., De Houwer, J., Pineno, O. & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 238 – 249.

Bouton, M. E. (1994). Context, ambiguity, and classical conditioning. *Current directions in psychological science*, *3*(2), 49-53.

Bouton, M. E., & Todd, T. P. (2014). A fundamental role for context in instrumental learning and extinction. *Behavioural processes*, *104*, 13-19.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58,* 313–323.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, *10*(7), 294-300.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *The Quarterly Journal of Experimental Psychology, 36*, 29-50.

Fiske, S. T., & Taylor, S. E. (1984). *Social cognition (2nd Ed.)*. Reading, MA: Addison-Wesley.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

Graham, C. H., & Gagné, R. M. (1940). The acquisition, extinction, and spontaneous recovery of a conditioned operant response. *Journal of Experimental Psychology, 26*(3), 251–280.

Guttman, N. (1953). Operant conditioning, extinction, and periodic reinforcement in relation to concentration of sucrose used as reinforcing agent. *Journal of Experimental Psychology, 46*(4), 213–224.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MA: MIT Press.

Haselgrove, M., & Evans, L. H. (2010). Variations in selective and nonselective prediction error with the negative dimension of schizotypy. *The Quarterly Journal of Experimental Psychology*, *63*, 1127-1149.

Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European journal of social psychology*, *20*, 311-335.

Herman, L. M. (2006). Intelligence and rational behaviour in the bottlenosed dolphin. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (p. 439–467). Oxford University Press.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual review of psychology, 53*, 575-604.

Hintze, J.L., & Nelson, R.D. (1998). Violin Plots: A Box Plot-Density Trace Synergism, *The American Statistician, 52*, 181-184

Holland, P. C., & Coldwell, S. F. (1993). Transfer of inhibitory stimulus control in operant feature-negative discriminations. *Learning and Motivation*, *24*(4), 345-375.

Holmes, N. M., Chan, Y. Y., & Westbrook, F. (2019) A combination of common and individual error terms is not needed to explain associative changes when cues with different training histories are conditioned in compound: A review of Rescorla's compound test procedure. *Journal of Experimental Psychology: Animal Learning and Cognition, 45,* 242-256.

Jeffreys, H. (1961). *The Theory of Probability (3rd Ed.)*. Oxford: Oxford University Press.

Jones, P. M., & Pearce, J. M. (2015). The fate of redundant cues: Further analysis of the redundancy effect. *Learning & Behavior 43*, 72-82.

Jones, P. M., & Zaksaite, T. (2018). The redundancy effect in human causal learning: no evidence for changes in selective attention. *Quarterly Journal of Experimental Psychology, 71*, 1748-1760.

Jones, P. M., Zaksaite, T., & Mitchell, C. J. (2019). Uncertainty and blocking in human causal learning. *Journal of Experimental Psychology: Animal Learning and Cognition, 45*, 111-124.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*, 34-45.

Kamin, L. J. (1969). Selective association and conditioning. In N. J. Mackintosh & W. K. Honig (Eds.), *Fundamental Issues in Associative Learning* (pp. 42–64). Halifax, Canada: Dalhousie University Press.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological review*, *113*(4), 677.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & behavior*, *36*(3), 210-226.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of mathematical psychology, 45*, 812-863.

Le Pelley, M. E., & McLaren, I. P. L. (2001). The mechanics of associative change. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology Section B, 57*, 193-243.

Lewis, D. J. (1956). Acquisition, extinction, and spontaneous recovery as a function of percentage of reinforcement and intertrial intervals. *Journal of Experimental Psychology, 51*(1), 45–53.

Levitt, J. T., Brown, T. A., Orsillo, S. M., & Barlow, D. H. (2004). The effects of acceptance versus suppression of emotion on subjective and psychophysiological response to carbon dioxide challenge in patients with panic disorder. *Behavior therapy*, *35*, 747-766.

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

Lubow, R. E. (1973). Latent inhibition. *Psychological bulletin*, *79*(6), 398.

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: the effect of nonreinforced pre-exposure to the conditional stimulus. *Journal of comparative and physiological psychology, 52*, 415-419.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276–298.

Marino, L., & Colvin, C. M. (2015). Thinking pigs: A comparative review of cognition, emotion, and personality in Sus domesticus. *International Journal of Comparative Psychology, 28*, Article 23859.

Miller, R. R., & Matute, H. (1996). Biological significance in forward and backward blocking: Resolution of a discrepancy between animal conditioning and human causal judgment. *Journal of Experimental Psychology: General*, *125*, 370-386.

Mitchell, P. (1997). *Introduction to theory of mind: Children, autism and apes*. Edward Arnold Publishers.

Mitchell, C. J., & Lovibond, P. F. (2002). Backward and forward blocking in human electrodermal conditioning: Blocking requires an assumption of outcome additivity. *The Quarterly Journal of Experimental Psychology*, *55*, 311-329.

Mitchell, C. J., Harris, J. A., Westbrook, R. F., & Griffiths, O. (2008). Changes in cue associability across training in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*, 423-436.

Moghadam, M., Ashayeri, H., Salavati, M., Sarafzadeh, J., Taghipoor, K. D., Saeedi, A., & Salehi, R. (2011). Reliability of center of pressure measures of postural stability in healthy older adults: effects of postural task difficulty and cognitive load. *Gait & posture*, *33*, 651-655.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & cognition*, *22*, 352-369.

Pavlov, I. P. (1927). *Conditioned reflexes* (G. V. Anrep, Trans.). Oxford: Oxford University Press.

Pearce, J. M., Dopson, J. C., Haselgrove, M., & Esber, G. R. (2012). The fate of redundant cues during blocking and a simple discrimination. *Journal of Experimental Psychology: Animal Behaviour Processes*, *38*, 167-179.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*, 532–552.

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8-13. www.psychopy.org.

R Core Team. (2018). *R: A language and environment for statistical computing*. www.r-project.org.

Rescorla, R. A. (1971). Summation and retardation tests of latent inhibition. *Journal of Comparative and Physiological Psychology, 75*(1), 77–81

Rescorla, R. A. (2000). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behavior Processes, 26*, 428–438.

Rescorla, R. A. (2001). Unequal associative changes when excitors and neutral stimuli are conditioned in compound. *Quarterly Journal of Experimental Psychology, 54B*, 53–68.

Rescorla, R. A., and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64-99). New York, NY: Appleton-Century-Crofts.

Seabrooke, T., Wills, A. J., Hogarth, L., & Mitchell, C. J. (2019). Automaticity and cognitive control: Effects of cognitive load on cue-controlled reward choice. *Quarterly Journal of Experimental Psychology*, *72*, 1507-1521.

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis.* New York: D. Appleton-Century.

Spicer, S. G., Mitchell, C. J., Wills, A. J., and Jones, P. M. (2019). Theory protection in associative learning: humans maintain certain beliefs in a manner that violates prediction error. *Journal of Experimental Psychology: Animal Learning and Cognition*.

Spicer, S. G., Mitchell, C. J., Wills, A. J., Blake. K. L., and Jones, P. M. (under review). Theory protection: do humans protect existing associative links? *Journal of Experimental Psychology: Animal Learning and Cognition*.

Spicer, S. G., Mitchell, C. J., Wills, A. J., Dome, L., and Jones, P. M. (under review). Representing uncertainty in the Rescorla-Wagner model: blocking, the redundancy effect, and outcome base rate. *Open Journal of Experimental Psychology and Neuroscience.*

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B, 37B*, 1-21.

Thorwart, A., Schultheis, H., König, S. & Lachnit, H. (2009). ALTSim: A MATLAB simulator for current associative learning theories. *Behavior Research Methods, 41*, 29-34

Tomasello, M., & Call, J. (1994). Social cognition of monkeys and apes. *American Journal of Physical Anthropology*, *37*, 273-305.

Uengoer, M., Dwyer, D. M., Koenig, S., & Pearce, J. M. (2019). A test for a difference in the associability of blocked and uninformative cues in human predictive learning. *Quarterly Journal of Experimental Psychology*, *72*, 222–237.

Uengoer, M., Lotz, A., Pearce, J. M., (2013). The fate of redundant cues in human predictive learning. *Journal of Experimental Psychology: Animal Behaviour Processes, 39*, 323-333.

Vandorpe, S., De Houwer, J., & Beckers, T. (2007). Outcome maximality and additivity training also influence cue competition in causal learning when learning involves many cues and events. *Quarterly Journal of Experimental Psychology, 60*, 356-368.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8*, 600-608.

Walsh, C. R., & Johnson-Laird, P. N. (2009). Changing your mind. *Memory & Cognition, 37*, 624-631.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.

Wills, A. J., Graham, S., Koh, Z., McLaren, I. P., & Rolland, M. D. (2011). Effects of concurrent load on feature-and rule-based generalization in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37*, 308.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological bulletin, 138*, 102-125.

Wills, A. J., O'Connell, G., Edmunds, C. E., & Inkster, A. B. (2017). Progress in Modeling Through Distributed Collaboration: Concepts, Tools and Category-Learning Examples. *In Psychology of Learning and Motivation* (Vol. 66, pp. 79-115). Academic Press.

Wills, A. J., Dome, L., Edmunds C. E., Honke, G., Inkster, A. B., Schlegelmilch, R., & Spicer, S. G. (2019). catlearn: Formal Psychological Models of Categorization and Learning. *R package version 0.6.2.*

Zaksaite, T., & Jones, P. M. (2019). The redundancy effect is related to a lack of conditioned inhibition: Evidence from a task in which excitation and inhibition are symmetrical. *Quarterly Journal of Experimental Psychology, 73,* 260-278.