
# Prediction error and cognitive control in human causal learning

Justine Kate Greenaway

School of Psychology
Faculty of Science
The University of Sydney

*A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy*
*2024*

**ABSTRACT**

Prediction error, the extent to which an outcome is unexpected, is thought to be an important determinant of the strength of learning. The blocking effect in human causal learning is consistent with this notion as it suggests that people fail to encode an association between a cue and outcome if that outcome is already expected on the basis of other predictive cues. In forming causal judgments, humans are clearly capable of drawing on other knowledge, such as information provided by testimony, or inferences derived from more complex reasoning processes like deduction. Nonetheless, it is plausible that associative memory processes and reasoning processes might both contribute to the expression of causal judgements in humans. If we take the assumption that both processes are involved, then this raises the question, to what extent do these processes operate independently—or alternatively, interact—to determine causal learning? This thesis examined one possible way in which these interactions may occur: that the predictions important for prediction error learning are subject to control by non-associative information. To that end, I provided instructions that manipulated the relevance of associative history to deriving predictions in a prediction error causal learning task. In Chapter 2, such instructions reduced the magnitude of the blocking effect in causal judgements, yet they did not appear to affect prediction error. That is, associative memory encoding was surprisingly resistant to instructed control. Chapter 3 examined the effect of providing a more explicit source of predictions which were either consistent with or conflicted with associative history. Finally, Chapter 4 investigated the influence of motivation on the efficacy of instructions that manipulate the relevance of associative history. The theoretical and practical implications of the findings in this thesis are discussed in the final chapter.

**ACKNOWLEDGEMENTS**

First and foremost, I would like to express my gratitude to my supervisor, Evan Livesey. This has in many ways been a rather humbling experience. Though my confidence was often shaken, I could always rely on your steadfast confidence in me. Your quiet encouragement got me through the weeds countless times. Thank you for being so generous with your time, guidance, and support over the years. You are a truly excellent supervisor. I also wish to thank my associate supervisors Justin Harris, Micah Goldwater, and Dan Pearson for their invaluable feedback and guidance.

To the amazing young researchers I met along the way, I am so grateful for your friendship. Thank you, Dot, your calm presence and pragmatism helped me feel at home in my doctorate. Thank you to Hil for being a wonderful mentor in many things, research, baking, confidence, and life. And to Julie, my PhD sister, thank you for the laughs, bouldering sessions, COVID walks, crosswords, and all the little moments I can't fit here. I am so excited to see what you all achieve. Thanks also to the Dogs, who helped me laugh at myself and taught me about all the big stuff like wine, plant care, and MarioKart8.

I also wish to thank my parents, who despite their playful ribbing about my being a 'permanent student,' have always been my greatest supporters, both in spirit and tangibly. I love you and owe you everything.

Lastly to my little family, Joel and Rue, thank you for being my raison d'etre. I love you endlessly, so I dedicate this thesis to you (even though you'll likely never read it).

**STATEMENT OF ORIGINALITY**

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

*Justine Greenaway, 08.01.24*

**TABLE OF CONTENTS**

## LIST OF TABLES



## LIST OF FIGURES

# CHAPTER 1

## General Introduction

Each day we make judgements that are grounded in our knowledge of the causal structure of our environment. The ability to infer causality from the statistics of our environment enables us to predict and control upcoming events to our advantage. For instance, noticing that you consistently suffer a stomach-ache after breakfast enables you to intervene to prevent future stomach aches by avoiding those foods causing the discomfort. Understanding the processes that enable us to learn from our environment then is an important goal for psychology. One candidate process is provided by accounts of associative learning. Broadly, these accounts attempt to explain how we acquire contingency information, the knowledge that events 'go together,' through repeated experience with the pairing of these events in our environment. The assumption of many theories of associative learning is that, all else being equal, this experience results in a link between the mental representations of these events, such that encountering one event will trigger retrieval of its associate. I will refer to this class of theory as *associative* in that they share a core theoretical construct—associative links— and are primarily concerned with the strengthening and weakening of those links. In this thesis I will refer to these links as *associative memories* and the process by which they automatically retrieve the outcome as *associative memory processes*.

There are several competing formal models that aim to capture how and when these associations are formed, but the general principles of associative theory have been applied to human causal learning with relative success (Shanks, 2007; Le Pelley et al., 2017; Thorwart & Livesey, 2016). Such accounts propose that a learned association between events may be interpreted as evidence that they are causally linked. For instance, if you repeatedly experience stomach ache after eating breakfast you may come to associate breakfast with developing a stomach ache. From this evidence, you may infer that your breakfast is *causing* the stomach ache and, as a result, change your breakfast habits.

Yet for many reasons, the problem of how to apply causal induction is more complicated (and the cognitive process by which we achieve it is more complex and sophisticated) than the previous example implies. Associative learning enables us to detect a statistical relationship between events that we experience, but those events need not be causally related. For instance, in a fear conditioning experiment, a rat will learn to fear a light that precedes a footshock, but the light does not cause the footshock, it merely signals the delivery of the shock. Similarly, the wearing of face masks is strongly associated with contagious illness, but it would be fallacious to believe that face masks cause illness. In other words, a correlation between events does not necessarily indicate a causal connection. As such, an associative memory may serve as a useful but limited heuristic for a causal relationship. Not surprisingly, people draw on other information to reason about cause and effect. For instance, we can read about a clinical trial that has shown that folate is important during pregnancy, or hear from a friend or a marketing representative that a new drug is effective in relieving pain, thus acquiring causal beliefs without ever having the ability to draw such a conclusion from personal experience. We can retrospectively list the foods that we have eaten for breakfast over many days and create hypotheses about which ingredients made us feel sick. We can engage in reasoning based on these hypotheses - if it were coffee that made me ill at breakfast then it should have also made me ill when I have my afternoon cup but instead I feel fine.

Associative theories, which often take the form of mathematical algorithms (e.g. Rescorla & Wagner, 1972), formalise one aspect of information accumulation very precisely. In doing so, however, they make the assumption that learning lawfully follows a predictable path; they do not directly speak to the influence of other cognitive processes on causal learning. Indeed, theorists often make the assumption that the association formation operations that these models describe proceed automatically (e.g. McLaren et al., 2014). In reality, there may be many situations in which cognitive processes interact as we come to understand cause and effect. To return to our example, your typical breakfast may consist of a number of candidate allergens but reading about gluten intolerance could change the inferences you make about which ingredients in your breakfast are

causing you to feel unwell. Indeed it is possible that many of the things you might learn and remember about these ingredients (not just their likely status as an allergen) are influenced by such knowledge.

In this chapter, I will evaluate the theoretical accounts for causal learning based on associative memory processes and higher order reasoning processes, and review the evidence for each. I will then propose that there is sufficient reason to assume that causal learning involves both of these processes. There is no doubt that we are capable of complex reasoning about causal relationships in a manner that moves beyond relying on simple associative memory retrieval. However, as will be discussed throughout this thesis there may be many circumstances in which we rely on intuitive judgements based on the ease with which an effect is brought to mind in the presence of a candidate cause. If causal judgements can be made on the basis of associative memory, and/or via reasoned inferences that draw on other processes, then we might ask whether and to what extent these processes influence one another. For example, how does information about the *relevance* of our experience with the candidate cause influence what we learn about it when we encounter it again? This thesis will examine interactions between processes that enable us to acquire and apply associative information to understand our environment.

**Associative accounts of causal learning**

The connection between association and causal knowledge was first made by the Empiricist philosopher Hume (1740/2003). Hume argued that as we are unable to observe causal forces directly, we must infer causality from our direct experience with the 'constant conjunction' between events – the statistical regularities in our environment that connect a cause to an effect. Experiencing these regularities will, Hume argues, lead one to develop an expectation that an effect of interest will follow when we encounter the cause. The expectancies we develop then serve as evidence when similar events are encountered in the future, thus, causal beliefs are derived from direct experience or inferred from past experience with similar events or objects. I will refer to the notion that causal beliefs are informed by associative memories as the *associationist* assumption.

This section will provide a brief review of the associationist approach to causal reasoning and it's theoretical development in human and animal learning.

Consistent with an empiricist view of learning, both humans and non-human animals are sensitive to the contingency between events in their environment and modify their actions on the basis of that experience. Contingency refers to the statistical relatedness of two events, and is defined formally in the ΔP model as the difference between the probability that the outcome will occur given the cue is present, and the probability that the outcome will occur in the absence of the cue ($p(O|C) - p(O|{\sim}C)$; Allan, 1980). In a particularly elegant set of experiments, Rescorla (1968, 1972) demonstrated that the strength of a conditioned response (CR) is related to the degree of contingency between a cue (or conditioned stimulus, CS) and an outcome (unconditioned stimulus, US). CS-US pairings alone were insufficient to produce conditioned fear in rats, with the animals only acquiring conditioned fear of the CS if the US was contingent upon it, that is if the US was more likely to occur in the presence of the CS than in the absence of the CS, even when they had experienced an equal number of CS-US pairings.

Similarly, early studies of contingency learning embedded in a causal scenario found that humans are able to detect with reasonable accuracy the contingency between a response and an outcome (Allan & Jenkins, 1983; Jenkins & Ward, 1965; Shanks & Dickinson, 1991; Wasserman et al., 1983) and between a cue and outcome (Shanks, 1991; Wasserman, 1990). Importantly, this is reflected not only in behavioural changes but also in judgements of causality. For instance, Shanks and Dickinson (1991) demonstrated that both the perception that an action caused a particular outcome, and the frequency of performing the action, varied systematically with the contingency between that action and the outcome. That is, causal judgements tend to accurately reflect the degree of contingency between the candidate cause and the effect, with some notable exceptions.[1]

[1] Causal judgements track contingency with reasonable accuracy with the exception of situations of zero contingency. In this instance, we tend to overestimate the contingency between two unrelated events when the frequency of either event (the cue or outcome) is inflated. Cue and outcome density effects in causal learning have been studied extensively (e.g. Matute et al., 2019). It is worth noting that these effects were

This is consistent with the notion that humans are able to track event contingency, and crucially, infer causality on the basis of that information.

Given this sensitivity to contingency, the obvious question is how contingency information is acquired? Coinciding with profound shifts in the focus of animal learning theory and research in the 1970s, researchers began explicitly noting the similarities between contingency learning phenomena in human cognitive experiments and the patterns of anticipatory behaviour widely observed in Pavlovian and instrumental conditioning. The seminal example of this used a video game scenario to investigate selective learning effects in instrumental causal learning. For reasons discussed in detail later, Dickinson et al. (1984) were able to demonstrate that participants' judgements of the contingency between an action (firing a gun) and outcome (destroying an enemy tank) were subject to the same phenomena seen in animal conditioning studies. These comparisons led some to posit that the psychological mechanisms underpinning conditioning are similar to those supporting human contingency learning, and thus to apply the same models of learning to both.

According to the associationist hypothesis, associative memories are developed through experience with covariation among cues and outcomes and, with this experience, the memory of an outcome is automatically retrieved in the presence of a predictive cue. It is proposed that the ease of recall of an outcome in the presence of a cue may serve as evidence of a causal relationship (Shanks, 2007; Le Pelley et al., 2017, Thorwart & Livesey, 2016). That is, the extent to which we predict an outcome given the presence of a cue and our decision about whether the cue *causes* the outcome are related in that both rely on associative retrieval.

The central tenet of many associative learning models is that the strength of learning is determined by prediction error, or the extent to which the outcome is surprising (e.g. Rescorla & Wagner, 1972). Such models assume that co-occurring cues compete for association with an outcome, and thus anticipate a range of widely observed behavioural phenomena known

---

anticipated by associative models of learning and have also been demonstrated in animal conditioning (Rescorla, 1972).

collectively as cue competition effects. However, as I will discuss, participants' predictions and causal judgements derived from cue-outcome relationships also reflect the influence of processes that are not captured by these models, such as inference based on contextual information provided by instructions.

**The blocking effect**

The blocking effect (Kamin, 1968) is one example of cue competition that was particularly important for the development of associative learning theory and its application to human causal learning. Blocking revealed a key principle of associative learning that went on to inform the most influential models of associative learning to date. Specifically, blocking demonstrated that contiguity and contingency between events alone is insufficient to produce learning, rather learning only occurs when the outcome is surprising. The first demonstration of blocking used a conditioned suppression paradigm in rats whereby conditioned fear of a stimulus was measured by the extent to which lever pressing behaviour was suppressed in the presence of that stimulus (Kamin, 1968). The design of this experiment is given in Table 1.1. In a pretraining phase, one group of animals (Group 1) received repeated presentations of a conditioned stimulus (a tone – 'A') followed by an electric shock to the foot (+). At the end of pretraining, the rats in Group 1 demonstrated conditioned fear of the tone, as evidenced by a reduction of lever pressing behaviour when the tone was presented.

In the next phase, Group 1 were presented with the tone along with another stimulus (a light – 'B'), and this compound was again followed by a footshock (AB+). Group 2 were given the same number of presentations of the compound and footshock (AB+) but did not receive the pretraining of the tone. Test trials in which the light was presented alone produced a significantly higher suppression ratio (less conditioned fear) to the light in Group 1 than in Group 2, despite receiving the same number of pairings of the light with the footshock. Kamin (1968) concluded that the pretraining of the tone blocked learning about the relationship between the light and the footshock because the light provided no new information for Group 1. The shock in the presence of AB was unsurprising as the outcome was already fully predicted by A, therefore B provided no new

information about the outcome. Blocking is one example of a group of effects collectively referred to as *cue competition* as they demonstrate that cues interact in learning, competing for association with the outcome.

**Table 1.1**

*Kamin's (1969) blocking design*

| Group | Pretraining | Training | Test |
|---|---|---|---|
| 1 | A+ | AB+ | B? |
| 2 | | AB+ | B? |

Blocking occurs not only in animal conditioning, but has also been demonstrated in human causal learning (e.g. Aiken et al. 2000; Livesey et al., 2019; Shanks, 1985). In causal learning research, the blocking design is embedded in a hypothetical scenario like the widely used *allergist task*, in which participants assume the role of a doctor diagnosing the food allergies of a fictitious patient by learning which foods produce allergic reactions in their patient. For example, participants may first learn that their patient suffers an allergic reaction when they eat apples (A+). In a second phase they see that the patient also suffers an allergic reaction when they eat apples and bananas (AB+) as well as when they eat carrots and dates (CD+). When asked to make a judgement about the extent to which each food causes an allergic reaction in their patient, participants tend to give lower causal ratings for bananas (B) than they do for either carrots or dates (C/D) even though they each received the same number of pairings with an allergic reaction (Aitken et al., 2000). The control cue in this case (C or D) is one that has been subject to *overshadowing* (Pavlov, 1927). Overshadowing occurs when two novel cues (C and D) compete with each other for association with the outcome. In this case, although they are associated with the outcome, the causal status of C is equally ambiguous to that of D because they have only occurred together.

***Associative accounts of the blocking effect***

Kamin's (1968) observation that learning is partly determined by the 'suprisingness' of the outcome was later formalised in the most influential model of associative learning to date, the Rescorla-Wagner model (Rescorla & Wagner, 1972). The model captures trial-based learning, or a change in associative strength, in the following equation:

$$\Delta V_X = \alpha\beta(\lambda - \Sigma V)$$

where $\Delta V_X$ represents the change in associative strength ($V$) for cue X on the current trial. The model posits that $\alpha$ and $\beta$, representing the salience of the cue (or CS) and outcome (or US) respectively, affect the rate of learning. The strength of learning is given by the error term in parentheses. This term represents prediction error, or the discrepancy between what is experienced ($\lambda$, or the total amount of associative strength that a given outcome can support) and what is expected on the basis of prior learning ($\Sigma V$, or the sum of the associative strength of all cues present on a trial). Less formally, the error reflects how surprising the outcome of a trial is given the predictive value of all cues presented.

The Rescorla-Wagner model provides a simple explanation of blocking. The assumption that predictions that inform prediction error are derived from the *summed* associative strength of the cues present enables the model to explain the blocking effect. To illustrate, repeated pairings of a pretrained cue and outcome (A+) produces gains in associative strength, such that by the end of pretraining, $V_A$ is close to asymptote (λ). In the training phase, the summed associative strength of A and B will be close to asymptote, as A is already strongly associated with the outcome. As a result there will be little prediction error and therefore little gain in associative strength, the presence of A blocks learning of the B-outcome association.

The Rescorla-Wagner model therefore describes blocking as a learning deficit, a failure to encode the relationship between a redundant cue and an outcome. A corollary of this assumption is that memory for the relationship between the blocked cue and the outcome should be relatively weak, and some human causal learning results are clearly consistent with this suggestion. For

instance, Mitchell and colleagues (2006) gave participants the typical blocking design embedded in an allergist task. At test, they were asked not only to assess the causal efficacy of each cue but also to recall the outcome with which it had been paired during training. They found that a blocking effect in causal judegments was accompanied by poor recall for the outcome that had been paired with the blocked cue (Experiments 1 & 2)[2]. Indeed, it appears these memory deficits occur in a manner consistent with a relatively fast and automatic memory retrieval process. Morís and colleagues (2014) embedded a blocking design within a simple word association task and then used an associative recognition priming test to assess the extent to which the cues could automatically retrieve their associated outcomes. In the priming test, cues from training were presented briefly and followed by a target word that was either consistent (an outcome that had followed the cue) or inconsistent (an outcome that was not paired with the cue) with what they had seen during training. Participants were asked to indicate as quickly and accurately as possible whether the target word shown was new or old. They found that responses were faster for consistent than inconsistent primes only in the control condition. In other words, cue-outcome memory from training facilitated retrieval of the relevant target for control cues, but not for blocked cues, indicating a failure to encode or retrieve the blocked cue-outcome association.

***Attentional accounts of the blocking effect***

A closely related set of models describes the learning deficit that produces blocking effects as a phenomenon of selective attention rather than a failure to update the associative strength of the cue-outcome relationship on compound blocking trials. The most prominent of these, the Mackintosh (1975) model, assumes that attention is modified by experience with the predictive value of cues. Cues that reliably predict a given outcome will receive increased attention and attention towards cues that are less predictive will decrease. In other words, according to this

[2] Although this may appear *prima facie* to be evidence of associative memory processes in human causal learning, the authors provided an alternative explanation for the results. They argued that participants inferred that the blocked cue was not causal *during training* and, as a result, did not encode the cue-outcome relationship (Mitchell et al., 2006). The specific inferences that may support blocking are discussed in the upcoming section on inferential accounts of causal learning.

account, prediction error produces changes in selective attention to cues, and it is these changes to attention that affect the rate of learning. If we return to our example of blocking from causal learning studies, during the pretraining, repeated experiences of A+ will increase attention to A as a reliable predictor of the outcome. The increase in attention to A will in turn increase the rate of learning such that the A+ association is acquired quickly. When AB+ is encountered in the training phase, the good predictor A will be well attended, and attention to the poorer predictor B will decrease leading to poorer learning of the B+ association. In other words, blocking is the result of a learned reduction in attention to the blocked cue during training. There are a number of alternative attentional models of learning that give similar explanations of blocking (weaker acquisition of the blocked cue-outcome association due to reduced attention to the blocked cue), and while they vary in how and when changes in attention are implemented in the model they all make an appeal to prediction error as the driver of attentional shifts (Kruschke, 2003; Le Pelley, 2004; Pearce & Hall, 1980).

Consistent with this account, there is empirical support for learned attentional shifts in blocking. Previous studies have found, for instance, that participants direct less overt attention (as measured by eye gaze) towards blocked cues than either pretrained or control cues (Beesley & Le Pelley, 2011; Kruschke et al., 2005; Wills et al., 2007). Indeed it appears that the blocking effect can produce changes in the extent to which a cue captures attention. Luque et al. (2018) instantiated blocking in an associative learning task and later presented the blocked and control stimuli as cues in a dot probe task. Responses were slower if the target was cued by a blocked cue than by a control cue, indicating that blocking produces a reduction in rapid attentional capture.

Models of attention in learning also propose that learned attentional biases can be captured in parameters within the prediction error model that affect the representation and associability of a cue that enters into learning (Mackintosh, 1975; Pearce & Hall, 1980). According to such models, the attention paid to a cue determines the extent to which it can enter into new associations (its associability). Again, there is evidence for reduced associability in blocking, as learning about the

blocked cue appears to be retarded in a subsequent task, and this retardation is accompanied by a reduction in overt attention to the blocked cue (Beesley & Le Pelley, 2011, Kruschke & Blair, 2000; Le Pelley, et al., 2007). In other words, blocking is associated with a reduction in the extent to which the blocked cue is processed. This indicates that blocking effects likely involve learned selective attention and that the attentional shifts observed are governed by prediction error.

**The role of prediction error**

Although they differ in the specifics, most contemporary associative learning models posit that learning is driven by an error correction process. Prediction error models of learning, like the Rescorla-Wagner model, have been relatively successful in anticipating and explaining cue competition effects in human contingency learning (see McLaren et al., 2014 for a review). The Rescorla-Wagner model is an early example of a *delta rule* model (Widrow & Hoff, 1960), in which learning weights are strengthened and weakened proportional to the difference between the predicted outcome and the observed outcome (that is, synonymous with summed prediction error, in associative learning models). Similar instantiations of delta rule learning have been widely used in back-propagation networks (e.g. McClelland & Rumelhart, 1985) and neural networks (see Enquist & Ghirlanda, 2005) to explain a wide variety of human and animal behaviour.

The concept of prediction error also has broad appeal for explaining phenomena across different domains of information processing. For instance it is incorporated in dominant theories of sensory processing and perception (Press et al., 2020), reward processing and reinforcement learning (Sutton & Barto, 1981), and whole brain function (Clark, 2013; Friston, 2010). The assumption that prediction error is a crucial component of learning is supported by neurological evidence. For instance, the brain appears to produce neural representations of prediction error, possibly via domain-general circuits responsible for sensory, social, cognitive and reward prediction errors (see Corlett et al., 2022 for a recent meta-analysis). Studies of reward learning in non-human animals have shown that dopaminergic activity in the midbrain not only correlates with prediction error but that this activity appears to conform to the general principles of prediction error learning

with a summed error term (see Nasser et al., 2017, for a review). For instance, Waelti et al. (2001) demonstrated a blocking effect in the behaviour of monkeys that was accompanied by a corresponding blocking effect in the dopamine response recorded in the midbrain. In a rat study, Steinberg et al. (2013) intervened on these networks causally, directly stimulating the dopamine neurons implicated in prediction error learning. They found that stimulating activity during the compound training of a blocking design can effectively mimic a *positive* prediction error (one in which the reward experienced is unexpected). This artificially-produced prediction error prevented blocking from occurring by driving learning about the cue that would typically be blocked.

What these error signals encode, the type of information they contain and the kind of learning they support, is a matter of debate. However, there is some evidence that these dopamine prediction errors go beyond simply encoding the value of the reward. For instance, specific patterns of firing recorded in a small number of dopamine neurons in rats have been linked to a sensory prediction error – one in which the identity but not the reward value of an outcome is unexpected (Takahashi et al., 2017; Stalnaker et al., 2019). Similar sensory prediction errors have been identified in fMRI recordings taken from the midbrain of humans (Stalnaker et al., 2019). Furthermore, dopamine error signals also appear to be critical to the formation of stimulus-stimulus associations in which no reward is presented (Sharpe et al., 2017).

There is some evidence that memory formation and consolidation is facilitated by prediction error. Dopaminergic signals are associated with plasticity in the hippocampus, a region implicated in memory encoding and consolidation (see for example Lisman & Grace, 2005). More directly, positive reward prediction errors have been linked to improved declarative (De Loof et al., 2018; Greve et al., 2017) and incidental (Jang et al., 2018, Stanek et al., 2019) memory encoding. Although the exact nature of the link between prediction error, dopamine, and memory remains unclear, there is reasonably compelling converging evidence for learning processes governed by prediction error across different domains of action, cognition and perception.

Based on the wide ranging evidence consistent with prediction error driven learning, and the ubiquitous nature of memory retrieval, I take the assumption that there is strong justification to hypothesise that a) retrieval of memories may provide a source of evidence for more explicit beliefs, judgments and decisions about cause and effect, and b) prediction error based learning mechanisms need to be regarded as a plausible causal mechanism for the laying down of these memories in the first place. Even if these assumptions are true, the precise role that prediction error learning might play in causal reasoning is, of course, open to debate and may vary depending on other mechanisms at play.

**Inferential accounts of causal learning**

Although we have established that prediction error is an important determinant of the strength of causal (and non-causal) learning, the associationist account is clearly not equipped to explain causal learning in all of its complexity. The output of an associative learning process is an unsophisticated form of evidence of a causal relationship, in the sense that it merely encodes the information that that two events are statistically related, not how or why they are related. Associative models are consequently limited by their failure to account for other important influences on human learning. For example, they cannot capture the influence of the causal structure of the task on causal judgements (Van Hamme et al., 1993; Waldmann, 2000, 2001; Waldmann & Holyoak, 1992), or the capacity to deliberately select items for attention (Mitchell et al., 2012).

Inferential accounts of human learning present an alternative to the associationist approach to causal learning that appeals to higher-order reasoning processes. These accounts propose that propositions, not associations, form the basis of human learning (De Houwer, 2009; Mitchell et al., 2009). That is, the information learned from the statistical regularities of our environment is propositional in nature – propositions are statements that contain information about how events are related e.g., 'A *causes* B' – and our causal judgements are the result of inferential processes made on the basis of these propositions. Single process accounts such as this attempt to provide a

model broad enough to encompass all human causal learning.

Due to the important position blocking has held in the development of learning theory, it is considered by some to be a litmus test for any model of learning. One prominent inferential account of the blocking effect that relies on deductive reasoning (henceforth the deductive inferential account) appeals to a relatively simple deduction about the causal status of the blocked cue. The explanation necessarily ascribes two simple assumptions about causality to the participants attempting to solve a blocking task. The first is that an effect can have two causes. The second is that causes have a linear additive effect; two causes when combined will increase the probability or the magnitude of the effect (Mitchell & Lovibond, 2002; Lovibond et al., 2003). The typical causal scenario in which the blocking design is embedded employs deterministic causal relationships – an allergenic food will always produce a reaction – along with a binary outcome – an allergic reaction either occurs or does not occur (e.g. Larkin et al. 1998). Under these circumstances participants learn that "A always leads to an allergic reaction" and "AB always leads to an allergic reaction of the same intensity or magnitude as A alone." If participants assume causes are additive, they may make the simple inference that as A is causal and the addition of B did not lead to a more intense allergic reaction, then B cannot be causal. In other words, participants may give the blocked cue a low causal rating because they deduce that, as the allergic reaction was not more severe when B was presented with a known cause, B could not be causal.

Evidence consistent with this explanation of blocking comes from studies that directly manipulate additivity assumptions by providing training and instructions about how causes combine in the given scenario (Beckers et al., 2005; Mitchell & Lovibond, 2002; Lovibond et al., 2003). For instance, Lovibond et al. (2003) gave explicit pretraining that demonstrated either the additive (I+, J +, IJ++, where ++ indicates a more intense allergic reaction) or non-additive (I+, J+, IJ+) causal structure of the task, alongside verbal instructions that described these rules. The additivity pretraining increased the magnitude of blocking, by reducing causal ratings for the blocked cue

relative to the non-additive condition. Indeed, De Houwer et al. (2002) argued that people generally assume additivity in such tasks and showed that demonstrating that the outcome is submaximal, or below ceiling in intensity, is sufficient to increase the strength of blocking in causal learning tasks.

Livesey et al. (2019) used an independent set of cues embedded in the learning task to measure assumptions of additivity. They found that non-additive pretraining successfully removed the assumption of outcome additivity, which was present when participants were not given explicit pretraining. However, they also found a persistent blocking effect following non-additive pretraining, which was equivalent in magnitude to the blocking effect following no pretraining at all. This is significant because it suggests that the additivity assumption *alone* is not sufficient to have a substantial influence on blocking. It was only in a group that received explicit pretraining that the outcome was additive that blocking significantly increased relative to the other groups. This increase in blocking was accompanied by an increase in self-reported confidence about judgments of the blocked cue, consistent with the notion that the participant was able to deduce that this cue was not causal.

Without additivity training or demonstrably submaximal outcomes, inferential blocking still may occur in the typical procedure because the causal status of B cannot be resolved (Cheng, 1997; Cheng & Holyoak, 1995; Waldmann & Holyoak, 1992). Participants may give B a moderately low causal rating because the inference that B is causal is prevented by the presence of the known cause A on blocking trials. That is, it may be the case that B is causal, but there is no opportunity to observe the effect of B alone, leading to uncertainty about whether or not B is causal. Note however, that the learner is similarly unable to resolve the causal status of the the overshadowed control cues C and D. Jones and colleagues (2019) argued that when additivity training has not been given, causal ratings reflect the participant's uncertainty about the status of the blocked cue. They demonstrated that confidence in the accuracy of causal judgements for the blocked cue was equivalent to that for the overshadowed control cues for which a clear inference is also impossible. Similarly, they showed that causal ratings are sensitive to manipulations of the base rate of the

outcome. In other words, participants' ratings for the blocked cue were affected by the probability of an allergic reaction in general; causal ratings were lower when the patient rarely suffered an allergic reaction than when an allergic reaction occurred often. This suggests that participants are using other evidence, in this case the base rate of events, in an attempt to reduce the ambiguity about the causal status of B.

Additivity assumptions can be applied not only to the magnitude of the outcome, but also to the probability of the outcome. Where the relationship between the putative cause and the effect is deterministic, as in the examples mentioned above, the deductions described by the inferential account can only be applied to effect magnitude. However, some early demonstrations of blocking in human causal learning employed probabilistic relationships, creating the right circumstances for applying the additivity rule to disambiguate the causal status of the blocked cue and thereby produce inferential blocking. For example, in Dickinson et al.'s (1984) video game task, the action of firing the gun did not destroy the tank on every trial, instead the action was related to the outcome probabilistically.

Similarly, the first demonstration of backward blocking (in which the the pretraining and training phases of the blocking design are given in reverse order) employed probabilistic relationships (Shanks, 1985). According to a deductive inferential account of blocking, it is not surprising that blocking is observed under circumstances in which the outcome is probabilistic in nature as the inference that the blocked cue is not causal is supported by the observation that the blocked cue, when combined with the pretrained cue, does not increase the probability of the outcome. Notably, the Rescorla-Wagner model (1972) does not anticipate backward blocking as associative strength for a cue can only be updated when that cue is present on a trial and therefore nothing new can be learned about B if the A+ training occurs after the AB+ training. The presence of retrospective revaluation effects like backward blocking in humans (and animals) has therefore been

used to discount associative explanations of causal learning[3]. However, backward blocking is not commonly observed in situations where the outcome probability or magnitude is not very obviously additive in nature (Livesey et al., 2019). Thus it is possible that whereas forward blocking receives contributions from both associative and other inferential processes, backward blocking is typically the result of the latter alone.

**Statistical models of causal induction**

Thus far I have discussed the history of, and evidence for, associative and inferential approaches to causal learning as two opposing explanations of causal induction. However, a third category of models exists which attempt to explain causal induction at the computational level. These models emerged from early experiments in contingency learning, which demonstrated that humans are capable of tracking contingency and making causal inferences on the basis of that information. One of the earliest was the delta p model (Allan, 1980) which provided a normative account of contingency learning. Delta p describes contingency as the result of a comparison of the probabilities of the critical events, in particular, the probability of the effect given the cause is present versus absent.

The progenitors of the delta p model (e.g., Allan, 1980; Allan & Jenkins, 1983) acknowledged that the model is a normative description of the problem of determining contingency and does not provide a process account of causal induction. That is, the model is not intended to capture the psychological processes by which one acquires contingency knowledge. Instead, they pointed to similarities between the results of their studies and the predictions of associative models of learning.

In contrast to the clear definition of delta p as a rational model of causal judgment (i.e., defining the problem one faces when judging cause and effect, *not* the processes one engages in to make such judgments), these early studies led to other statistical models of learning that did

[3] There are a number of associative explanations of retrospective effects like backward blocking, which assume associatively retrieved representations can be learned about when they are retrieved by other cues (e.g., Aitken et al., 2001; Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994). However, given that this thesis is not directly concerned with retrospective revaluation effects, I will limit my discussion of these here.

attempt to provide some psychological description of causal induction. Most notable among these is perhaps is the probabilistic contrast (PC) model (Cheng & Holyoak, 1995, Cheng & Novick, 1992).[4] The model was developed after the discovery of blocking effects in causal learning and the notable failure of the delta p model to account for blocking. The PC model is a modified version of the delta p model which takes the assumption that people infer causality by performing a mental comparison of the probability of the outcome in the presence and absence of a potential cause. Critically, when more than one candidate cause is present, the PC model proposes that the learner calculates these probabilities across a 'focal set' of events in which all other putative causes are held constant. Accordingly, the model can capture many effects that were taken to support associative accounts of causal learning. Most relevant to this thesis is the model's account of the blocking effect. When calculating delta p for the blocked cue (B), the focal set of events are those in which another potential cause is always present (the pretrained cue A). Judgements about the blocked cue are based on a comparison between the probability of the outcome in the presence or absence of the blocked cue, in circumstances where the alternative cause is always present. As the outcome occurs with equal probability on all trials where A is present, regardless of whether B is present or absent, delta p will be zero (indicating a null contingency between the blocked cue and the outcome) and participants consequently should give the blocked cue a low causal rating.

The inferential contrasts that the PC model describes are made at the point of causal judgement, so the model is not constrained by the order in which the statistical information is delivered. As a result, the PC model has no difficulty accounting for the presence of backward blocking in human causal learning (Shanks, 1985). Nor does it matter the format of that information, for instance people are able to make accurate causal judgements when covariation information is

[4] Cheng (1997) later proposed a modified version of the PC model - the power PC model - which describes causal induction as the result of logical deduction from prior assumptions about the *causal power* of objects or events and observable statistical regularities between events. The power PC model was explicitly intended to be a theory concerning psychological mechanism, and as such was a direct competitor for other process models. See Allan (2003), or Lober & Shanks (2000) for a critique of the power PC model. Since the model does not make predictions about memory and retrieval mechanisms (and their relationship with causal judgements), the studies in this thesis are beyond the scope of the model.

presented in summary form (Ward & Jenkins, 1965). This is in contrast to associative models like Rescorla-Wagner that describe the process of acquiring covariation information on a trial-by-trial basis in which the order of presentation is critical to the resulting associative strength. Indeed, statistical models like the PC model can account for the effect information that is not captured in associative models. One prominent example is the effect of causal direction. Waldmann & Holyoak (1992) for instance proposed that the causal model participants entertain affects the expression of cue competition in causal judegments. In their study, two groups were trained in a classic blocking design differing only in the cover story provided for the task. One group was given the typical allergist scenario in which cues were described as potential allergens (the predictive scenario). The other group was given a diagnostic scenario in which cues were described as potential symptoms of an illness (the outcome). In the predictive scenario, the blocked cue (B) is redundant as it is only ever presented with a cue already established as a cause (the pretrained cue A). However, in the diagnostic scenario, there is no reason to assume that symptoms of a disease are in competition with one another, many diseases produce multiple symptoms, indeed sometimes the particular combination of symptoms can be more diagnostic than a single symptom alone. Consistent with this, only the predictive group demonstrated blocking in their causal judgements. Although causal model effects have been replicated in blocking and other cue competition paradigms like overshadowing (Blanco et al., 2014; Waldmann 2000; 2001), there are some notable failures to observe such effects (e.g. Arcediano et al., 2005; Shanks & López, 1996). Similar to the effect of manipulating assumptions of additivity discussed above, some researchers have argued that causal model effects only occur when the task instructions emphasised the relevance of causal structure information (López et al., 2005).

Noting the consistency between assessments of causality measured in causal learning tasks and the output of Bayes' theorem (which formalises a method of updating prior beliefs on the basis of new observations), some theorists have noted that causal reasoning often closely approximates a Bayesian statistical reasoning process (e.g. Tenenbaum & Griffiths 2003). This approach has been

formalised in Bayesian accounts that, like the delta p rule, are designed to define the computational problem that causal learning mechanisms have evolved to solve. The causal support model (Griffiths & Tenenbaum, 2005), for instance, addresses how we might judge the likelihood that a causal link exists between two events given a set of observations about their co-occurence (a judgement about causal structure rather than causal strength). It provides a means of assessing the certainty with which one might hold a belief about a causal relationship, which Griffiths and Tenenbaum (2005) argued may influence judgments about the strength of the potential relationship under some circumstances. While a full exposition of this and other Bayesian inferential accounts is beyond the scope of this thesis, as a formalisation of the rational problem the learner faces when engaging in causal reasoning, the Bayesian inferential approach has been very influential (e.g. see Holyoak & Cheng, 2011).

Applying Bayes' rule to the standard blocking design in causal learning, Livesey et al. (2013) argued that blocking is Bayes rational but noted that it was unlikely people were explicitly reasoning in this way given that people find explicit reasoning with conditional probabilities very difficult. Indeed, when the blocking design was framed as a problem of conditional probabilities, very few participants in their study were able to respond correctly even though the majority of participants demonstrated blocking in their causal judgements.

**Models of psychological process as explanations for causal learning**

When considering models of causal learning and reasoning, there are key differences to bear in mind in terms of the aims, the scope, and the focus of individual theories. Among *process* theories, those concerned with the psychological mechanisms that contribute to causal cognition (including associative and propositional theories of learning), there are of course many differences in the psychological mechanisms that are invoked to explain a phenomenon like blocking. However, despite their differences, process theories share in common several properties that set them aside from *rational* theories, like the Bayesian framework suggested by Griffiths, Tenenbaum and colleagues. First, unlike rational theories, which attempt to provide a theoretical framework for

understanding the problem of causal inference, process theories are concerned with modeling specific cognitive mechanisms. They are principally concerned with capturing the qualities of a psychological process that might explain how we arrive at our causal judgments in the way that we do. In contrast, rational models are more concerned with the solution that we *should* arrive at if we are acting as rational learners and decision makers. Second, whereas rational theories, by their very nature, are designed to be unifying frameworks, there is no logical requirement for process models to satisfy this aim. Rational models attempt to explain many, if not all, phenomena in their sphere (in this case, causal reasoning in the broadest sense), under a particular formal approach. Process models, on the other hand, are designed to model a particular psychological mechanism that may (but also may not) be responsible for a range of relevant phenomena. The aim of the model, it can be argued, is to better understand the psychological process. If the model manages to capture a wide range of phenomena within a particular domain, then this might lend support to the idea that the process is deeply involved in that domain, but this is secondary to what the model should achieve. When it comes to psychological process, for instance, so long as it is plausible that several processes may co-determine our thoughts and actions then a model of any single process need not be unifying nor universally applicable to all phenomena within that domain.

Why is this important? In human learning, there is sufficient evidence that both associative memory and reasoning processes exist and affect thoughts and behaviour in *some* way. While both are processes that need to be understood, it is not necessary that either one fully accounts for patterns of causal judgment data produced by people. Arguably, a model of either process in isolation may never provide an accurate account of human causal learning. Entertaining theories of causal learning that assume these mechanisms are jointly responsible for causal judgements is thus justified. However, doing so raises other problems. Naturally, when one tries to combine models of different psychological processes, there is a risk of redundancy. In the case of blocking for instance, some process models attribute the effect to the process of acquiring causal information while others attribute it to logically derived judgements made once that information is acquired. The fact that

either one of these mechanisms might produce blocking seems like an unparsimonious theoretical outcome. However, parsimony *per se* is not a compelling reason to discard combined theories of this nature. For starters, the evidence regarding blocking suggests no single process provides a satisfactory account of the data available. But perhaps more importantly, the assumption does not lack parsimony in any meaningful way because there is clear evidence that we possess and readily apply competencies for associative memory and for symbolic reasoning. If we possess and frequently use both then why would they not both be relevant to such a fundamentally important activity as causal learning? Studying *how* these processes interact and jointly give rise to causal judgements could arguably be far more revealing than attempting to explain all phenomena with either process in isolation. The next section outlines the way in which I have sought to do this.

**Studying the interaction of associative and inferential processes**

The question of whether causal induction is underpinned by associative processes or propositional reasoning is an important one, and one that has generated much research. However, in this thesis, the debate between associative and inferential accounts of human causal learning is set aside. Specifically, I take the view that there is sufficient empirical evidence and theoretical plausibility to assume that both play some role in human learning. Evidence that cue competition in causal learning can be driven by propositional processing is not sufficient to demonstrate that associative processes are never involved in making causal judgements. Acknowledgment that associative memory and reasoning processes may jointly determine causal judgments allows us to begin asking how these processes might interact.

To an extent we can be agnostic about the specific instantiation of the memory process (or processes) that may produce effects like blocking in causal learning, although as the evidence reviewed above suggests, prediction error is a critical element. Le Pelley et al. (2017) note that there is now good evidence to suggest that human contingency learning is often consistent with the general principles of associative models if not always consistent with the specific instantiations of the models. Associative learning models usually assume that learning is lawful, following formal

algorithms that make no reference to higher-level thought processes, in a way that strongly implies associative learning should be blind to whatever cognitive processes are occurring at other levels of information processing. Theorists have thus argued that one can accept the importance of inferential reasoning for causal learning and still hold the view that associative learning retains some of its independent characteristics such that its operations are partly autonomous, distinct from controlled cognition (McLaren et al., 2014; Sloman, 1996). I will argue that this assumption remains largely untested.

**Controlled cognition and prediction error**

If prediction error is a determinant of the strength of learning then the question arises, what kind of predictions, or expectancies, are important for learning? If one accepts that controlled cognition plays an important role in causal learning, then outcome predictions and causal judgements may be based on a range of independent factors. They may be related to associative memory processes but, at the same time, should be dissociable from retrieved associations, depending on the circumstances in which the predictions and judgments are made. Predictions and causal judgments may also be dissociable from one another for the same reasons.

As an example, overt predictions, which are often collected on a trial-by-trial basis throughout learning, should not necessarily be based solely on the associative history of the cues present. To illustrate, imagine a participant in the allergist task had learned that their patient reliably suffers an allergic reaction whenever they eat apples, such that the presentation of apples strongly retrieves a memory of allergic reaction. If they were then instructed that their patient was given a drug to suppress allergic reactions, presumably they would predict that any previously observed allergenic food would no longer produce an allergic reaction. This shift should occur instantly, it does not require any further observation, suggesting it is not dependent on direct experience. The next time the patient eats apples the participant would again retrieve the memory of allergic reaction. However, the new information about the drug should cause them to revise their predictions about the likelihood that their patient will suffer an allergic reaction. Our overt predictions, then, may

reflect the history of the cue (its relationship with allergy) in the absence of other information, but we can also use non-associative information to make a reasoned prediction about the (reduced) likelihood of the patient suffering an allergic reaction. Assuming our predictions can be modified flexibly in this way, how does such a change in prediction affect further learning? In other words, if instructions can control the overt predictions we make, can they also control how associative memories are laid down by modifying the predictions that enter into prediction error learning? If at this point the patient did not suffer from an allergic reaction, it would not be surprising given the instructed knowledge of the drug. However, this outcome would still deviate from a prediction based on previous experience alone.

Prediction error models invoke one type of outcome expectancy based on retrieval of previous learning. Under some circumstances, this expectancy appears to be dissociable from explicit and reasoned expectancies. For example, in the Perruchet effect participants' explicit expectancies tend to follow a pattern of negative recency akin to a gambler's fallacy whereas anticipatory responses appear to follow a pattern of positive recency, indicative of learned expectancies derived from the associative strength of the cue on a given trial (Barrett & Livesey, 2010; Lee Cheong Lem et al. 2018; Perruchet, 1985). Some argue that these two forms of expectancy are so different that they are necessarily the products of two independent systems (Sloman, 1996; Squire, 1992). To date, there has been no systematic investigation of how these two forms of expectancy might interact to produce cue competition effects in causal learning and other explicit learning tasks. Thorwart and Livesey (2016) suggest that if an error driven memory process is involved in making causal judgements, then the operations of this process could be influenced by external factors (e.g. instructions) in three ways; by changing the stimulus inputs on the basis of which predictions are made, by modifying the prediction itself, thereby influencing prediction error more directly, or by changing the way retrieval translates to behaviour. In this thesis I aimed to test the second of these possibilities.

**Measuring associative memory and causal judgements**

In causal learning tasks, judgements usually take the form of a causal rating. The learning and decision processes that form the basis of these ratings remain unclear. As discussed, a common assumption (see Le Pelley et al., 2017; Thorwart & Livesey, 2016) is that associative retrieval informs causal judgements but is not their only determinant. Variation in causal ratings across different conditions will follow associative strength under some conditions, but may deviate from it when circumstances suggest that ratings need to be made in other ways. Consistent with this, causal judgements appear to be sensitive to instructions that provide some information about causal structure. For example, as previously discussed, instructions that indicate how cues combine to cause an outcome (i.e., additively or non-additively) can affect the magnitude of cue competition in causal judgements (Lovibond et al., 2003; Mitchell & Lovibond, 2002; Livesey, et al., 2019). It remains to be seen whether encoding of associative memory is itself sensitive to instructions in a similar manner.

For this reason, in all of the experiments in this thesis, I assessed associative retrieval and causal judgement separately with two distinct test questions. This procedure has been employed in previous research in an attempt to distinguish between competing accounts of causal learning (Don & Livesey, 2015; Mitchell et al., 2005, 2006; Shone et al., 2015). However, the focus of the experiments that follow is not to weigh in on debates about whether associative or inferential (or statistical) models of causal learning provide a more accurate account of the data. Instead the aims were motivated by considering the possibility that associative memory and reasoned inferences combine to produce causal cognition. If one takes this assumption as a theoretical starting point then it is important to ask whether and how they interact. For instance, this thesis is concerned with the latent expectancies that are thought to constrain associative memory (by producing predictions and prediction error). In particular, even though these expectancies are thought to be associatively generated, are they also modifiable by reasoned assumptions produced via simple instructions? This hypothesis inspired a set of related research questions with one broad aim; to determine whether

associative memory judgements (as measured by test questions that specifically probe cue-outcome recall) are subject to modification (or control) by instructions, in the same way as causal judgements.

**Chapter summary**

Chapter 2 examines the impact of instructed relevance of pretraining to the expression of the blocking effect in both learning scores and causal ratings. Three experiments introduce a change in patient between pretraining and training to reduce the relevance of the pretraining to predictions made during training. In these experiments, participants first learn about several unambiguous relationships between certain foods and allergic reactions in a pretraining phase, before being introduced to more cues in the learning phase, some of which (e.g., blocked and overshadowed cues) are presented in compound such that their individual causal status is ambiguous. Participants were then asked to make causal and associative memory judgments about each cue individually. Participants in the *same patient* conditions received causal learning embedded within a food allergist task in which the learner investigates the causes of a single patient's allergic reactions. Under the circumstances invoked in this scenario, there are strong reasons to consider the information learned during pretraining as being relevant to later learning about new and potentially ambiguous cues. Consequently, and consistent with previous literature (e.g. Mitchell et al., 2006), I expected to see evidence of blocking on both test measures. In contrast, participants in the *new patient* conditions were told after pretraining that all subsequent trials in the training phase referred to a new patient with different allergies to the first. In this context, information acquired during the pretraining phase should be far less relevant to judgments made about the ambiguous cues encountered in the training phase (i.e., eaten only by the new patient). The question of interest is whether this pretrained information still blocks learning about an added cue even under these conditions.

To anticipate, I found evidence of sensitivity to the instructed patient change in training predictions and causal ratings, however, there was a striking lack of sensitivity in learning scores at test. This is surprising as it suggests that the predictions important for learning are largely

determined by the associative history of the cue, even when that history may not be wholly relevant.

In Chapter 3 I attempted to push the relevance manipulation of chapter two further by giving directed and explicit instructions about the contingencies in training. Instructions about the new patient undermine the relevance of pretraining and thus increase uncertainty. In a sense, using these instructions may seem counterproductive to the learner (even though continuing to use the pretraining information may be nonsensical). In contrast, the direct instructions provided in Chapter 3 introduce new information that reduces causal ambiguity. To implement this, participants were instructed that the new patient was allergic to specific foods, which were then used as competing cues paired with the new cues of interest. The results revealed *some* evidence that more directed instructions were sufficient to overcome the insensitivity in learning observed in Chapter two, although it remains ambiguous whether they successfully reduced the blocking effect. Nevertheless, providing very explicit instructions regarding which cues are likely to be the causes of allergies in a new patient resulted in a reduction in learning about the competing cue. These experiments thus revealed evidence of an instructed blocking effect, where learning is impaired for cues that are in direct competition with a cue that is explicitly labelled as causal.

Chapter 4 examines the implications of increasing the task relevance of the instruction manipulation, that is, the relevance of the instructions to making accurate predictions during training. The intention of this inquiry was to establish whether the persistence of blocking observed in Chapter 2 could be attributed to participants' approach to the task. That is, if the participant's goal is to make accurate predictions during training, rather than to learn the relationships in the task, this could produce blocking in memory retrieval that is resistant to instructed control. I found that increasing the alignment between the instructions and the goal, either by describing the pretraining phase as a practice example (Experiments 4.1 and 4.2) or by removing the requirement to make active predictions during training (Experiments 4.3 & 4.4), led to greater sensitivity to the

instructions in learning scores. The results of these experiments as well as their implications for theories of human causal learning are discussed in Chapter 5.

# CHAPTER 2

## Instructed control of the blocking effect

In the present chapter I aimed to modify prediction error in a causal learning scenario with instructions that manipulate the relevance of associative history. To that end, I conducted three experiments that used instructions to manipulate the relevance of pretraining in a blocking design. The experiments that follow used the classic allergist task (described in Chapter 1) in which participants are given a hypothetical scenario of a doctor trying to discover the allergies of their patients. The fictitious patient consumes various foods across a number of trials and the participant is asked to predict the reaction the patient will have to the food. They receive feedback on their predictions and learn through trial and error which foods cause which outcomes in their patients. At the end they are tested on their memory for the food allergies of the patient and are asked to judge whether each food cue causes the observed outcome. The following experiments have three stages, a pretraining phase, a training phase, and a test phase. Explicit instructions were introduced after pretraining to modify the relevance of the learning that had just taken place. Specifically, participants were given instructions that indicated that the next phase comprised observations from either the same patient or a new patient.

This explicit instruction between the pretraining and training phases should alter participants' overt predictions about the relevant contingencies in the training phase. I hypothesised (and later confirmed in Experiment 2.3) that participants would presume that what they have learned about the first patient's allergies will have substantially less bearing on what they are about to learn about the new patient. For instance, if the learner had observed that one patient reliably suffers an allergic reaction when they eat apples, that information would be relevant to making a judgement about the allergies of that same patient if they suffer an allergic reaction whenever they eat a meal of apples *and* bananas. If instead they learned in the training stage that a *new* patient suffers an allergic reaction whenever they eat a meal of apples and bananas, the associative history of apples should be less relevant to the learner's predictions and judgements about the added cue

bananas. Therefore, after receiving the new patient instructions, predictions for the blocking compounds AB and CD should be more similar to those for the novel overshadowed compounds EF.

In the test phase, associative memory and causal judgements regarding the relevant contingencies were assessed separately. Participants were asked to recall the specific outcome paired with each cue (associative memory), and subsequently asked to judge the likelihood that the cue caused any kind of allergic reaction (causal judgement). I expected explicit instructions to have a strong effect on the causal inferences people make in such a task. The key question was whether these instructions would also influence *associative memory* for the cue-outcome pairs introduced in the training phase. If associative learning is driven by prediction error and those predictions can be influenced by cognitive manipulations in the same way as explicit causal judgements, then we would expect to see instruction-based modulation of cue competition effects in both associative memory and causal ratings. That is, informing participants that the examples in the training phase were from a new patient should weaken blocking in an equivalent fashion for both dependent variables. Alternatively, if prediction error learning is a function of associative predictions that are not sensitive to other information, then we should see a dissociation between associative memory and explicit causal judgements. That is, the instructions should affect the magnitude of cue competition in causal ratings, but associative memory should show blocking regardless of instructions. I tested these hypotheses in Experiment 2.1, replicated the results with the addition of eye-tracking in Experiment 2.2, and because there was an indication that these experiments were under-powered, replicated the results with a large sample in Experiments 2.3a and b.

## Experiment 2.1

Experiment 2.1 represents the first attempt to modify cue competition with an instruction that manipulates the relevance of pretrained associations. The design of Experiment 2.1 is outlined in Table 2.1. A set of blocking and overshadowing contingencies was established for each of two possible allergic reaction outcomes. I also included another set of contingencies in which one cue is first shown not to lead to allergic reaction in pretraining, before a combination of this and a novel

cue are shown to lead to one of the allergic reaction outcomes in the training phase. These contingencies are associated with a reduction in the overshadowing observed to the novel cue (Carr, 1974) and the effect is sometimes referred to as *reduced overshadowing* or *unovershadowing* (I will use this latter term henceforth). While unovershadowing is a cue competition phenomenon that has attracted interest for its own reasons (see for example Carr, 1974; De Houwer et al., 2002; Vandorpe & De Houwer, 2005; Wheeler & Miller, 2007), the simplest prediction error models of learning (e.g. Rescorla & Wagner, 1972) do not predict this phenomenon and other associative learning models do so by appealing to a range of alternative mechanisms (e.g. McLaren & Mackintosh, 2002; Thorwart et al., 2012; Wagner, 1981) that provide parameter-specific predictions. Interpreting differences in the strength of unovershadowing across measures is therefore complex and I include unovershadowing contingencies in these experiments mainly to balance out the design along with additional filler cues. These cues are not the main focus of the experiments and learning about these cues will not be reported or discussed in detail in this thesis.

The critical group manipulation of this experiment (and those that follow) concerned instructions delivered after completion of the pretraining stage and before commencement of training. The *same patient* condition were told that they would see more meals from the same patient in the next phase. This condition is similar to most of previous blocking experiments using the food allergist task. Although not all studies include a break or instructions between learning stages, the instructions delivered to this group should not have a strong influence on what participants in this group learn during subsequent training. The *new patient* condition, on the other hand, were told that they would be assessing a different patient in the training stage. Although there was no explicit reference to any of the previously causal cues specifically, the implication of these instructions is that what was learned during pretraining is no longer directly relevant.

Based on a wealth of prior studies that use the food allergist task to demonstrate blocking (e.g., Aitkin et al., 2000; Jones et al., 2019; Livesey et al., 2019; Mitchell et al., 2006), we would expect to see evidence of blocking in the form of lower causal ratings *and* poorer associative

memory for the blocked cues (B and D) relative to overshadowed controls (E, F, G and H), in the same patient group. If the new patient instructions successfully modify the predictions over which learning proceeds, blocking should be attenuated in the new patient group. If, on the other hand, the predictions important for learning are derived exclusively from direct experience with the cue and are thus insensitive to instructions that undermine the relevance of that experience, then we should see poorer retrieval of the outcome paired with the blocked cue regardless of how the instructions qualify the relationship between the pretraining and training phases.

**Table 2.1**

*Design of Experiment 2.1*

| | Pretraining | Training | Test |
|---|---|---|---|
| Blocking | A – O1 | AB – O1 | A, B |
| | C – O2 | CD – O2 | C, D |
| Overshadowing | | EF – O1 | E, F |
| | | GH – O2 | G, H |
| Unovershadowing | I – no O | IJ – O1 | I, J |
| | K – no O | KL – O2 | K, L |
| Fillers | | MN – no O | M, N |
| | OP – O1 | U – no O | U |
| | QR – O2 | V – O1 | V |
| | ST – no O | W – O2 | W |
| | | X – no O | X |
| | | YZ – no O | Y, Z |

*Note*. Letters represent cues, randomly assigned to different food items. O1 and O2 represent different allergic reactions ('*Nausea'* or '*Fever'*); 'no O' represents '*no allergic reaction*.'

## Method

### Participants

Sixty-eight undergraduate psychology students from the University of Sydney participated in exchange for partial course credit. A learning criterion of 60% accuracy during training was chosen to remove participants who did not show reliable evidence of learning to predict the outcomes. Three participants were excluded for failing to meet this learning criterion while one additional participant was excluded on the grounds of having previously completed a similar cue competition task. The remaining sixty-four participants (55 female, *M* age = 21.5) were randomly assigned to receive either the same patient (*n* = 32) or new patient (*n* = 32) instructions.

### Apparatus & Stimuli

The allergist task was programmed in MATLAB using the Psychophysics Toolbox extensions (Kleiner et al., 2007). The experimental stimuli were 300 x 300 pixel photographic images of foods (*avocado, banana, milk, carrots, corn, apple, garlic, peanuts, rice, pasta, cheese, fish, bread, eggs, mushrooms, chicken, coffee, lemon, strawberries, chocolate, cherries, olive oil, bacon, pineapple, prawns,* and *peaches*) accompanied by a text label. Two distinct allergic reactions, *fever* or *nausea*, as well as *no allergic reaction* served as the outcomes. For each participant, the food stimuli and allergic reactions were randomly allocated to cues (A-Z) and to outcomes (O1/O2) respectively.

### Procedure

Participants assumed the role of an allergist attempting to diagnose the allergies of a fictitious patient ('Miss Y' or 'Mr X'). On each trial, participants were presented with one or two foods that the patient had eaten and were asked to predict which of the two possible allergic reactions (nausea, fever, or no allergic reaction) would result by clicking the appropriate label. Corrective feedback was given after each response. Pretraining involved seven trial types listed in Table 2.1 presented in a randomised order twice per block over five blocks.

After pretraining was completed, the critical instructions were delivered. Participants in the same patient group were told that they would see more meals from the same patient in the next

phase. The new patient group were instead instructed that they would be assessing another patient in the next phase, without referring to any relationships between the allergies suffered by the two patients (a script of all task instructions is given in Appendix A). The training phase then proceeded in the same way as pretraining. Participants completed three blocks of twelve trial types shown in Table 2.1, presented twice per block in randomised order.

In the test phase, participants were asked to make two separate judgments about each of the cues from the training phase. The first assessed memory for the cue-outcome pairings, the second asked participants to make a judgment about the causal status of the cue. The difference between the memory and causal ratings was emphasised with the following instructions:

> *First, indicate which allergic reaction followed after Miss Y ate this food. This is a bit like a memory test to see how well you can recall which foods and reactions were presented together. Second, indicate to what degree the food caused an allergic reaction of any kind in Miss Y.*

Given that it is possible to make a causal judgement about a cue in the absence of a strong memory for the particular outcome with which it was paired, the causal rating was deliberately less specific than the memory test.

Cues were presented individually in randomised order and participants made a set of four ratings for each one. First, participants were asked to recall which of the outcomes followed the cue in question during training. Three linear analogue scales were presented, one for each outcome presented during training (*no allergy*, *nausea* and *fever*). Each scale ranged from “Never followed this food” to “Always followed this food.” Participants selected the point on each scale representing the extent to which the outcome followed the food in question during the training phase. Once all three ratings were completed, they proceeded to a causal rating. Participants were asked to specify the extent to which they believe the cue caused any kind of allergic reaction. This was performed on a similar scale ranging from “Definitely DOES NOT cause a reaction” to “Definitely DOES cause a reaction.” Screenshots of the task are available in Appendix B.

**Analysis**

Analyses for the pretraining and training phases had two aims, the first was to confirm that participants learned the relevant contingencies as the hypotheses regarding associative memory and causal ratings were made on this assumption. The second aim was to assess the immediate effect of the instructions on overt predictions made during training. Accordingly, we conducted planned analyses of accuracy for the critical compounds in the first block of training, after the instructions were delivered.

In the test phase, judgements of the strength of the association between the critical cues (blocked and overshadowed control) and the specific allergic reaction with which they were paired during training were of greatest interest. As such, raw memory ratings were converted to learning scores, calculated as the rating with respect to the correct outcome minus the mean of the ratings given to the incorrect outcomes.

[5] Learning scores could range from -100 to 100, where a score of -100 reflects a strong memory for the incorrect outcome, and a score of 100, a strong memory for the correct outcome. All measures, including the learning scores, causal ratings, and training accuracy were collapsed across trials of the same type (for example, learning scores for the blocked cue comprise the mean of the learning scores for cues B and D).

In this thesis, the primary analyses consisted of JZS Bayesian ANOVAs (hereafter referred to as 'Bayesian ANOVAs') with default prior options for the effects (r = 0.5 for the fixed effects). Bayes factors for the main effects indicate the likelihood of the data given the main effects model relative to a null model derived from the grand mean (denoted $BF_{10}$), or alternatively given the null model over the main effect model ($BF_{01}$). Accordingly, a $BF_{10}$ of two indicates the data is two times as likely under the alternative than the null model. For interaction terms, I report Bayes Factor Inclusion/Exclusion Across Matched Models, denoted $BF_{Incl}$ or $BF_{Excl}$ (Rouder et al., 2017). This

[5] Calculating scores in this way has been employed in past studies when memory for the specific outcome with which a cue has been paired, rather than memory for an allergic reaction in general, is of primary interest (e.g. Mitchell et al, 2012).

provides an estimate of evidence for or against an interaction by comparing the models that contain the interaction effect against equivalent models stripped of the interaction effect. Although Bayes Factors enable a more nuanced interpretation of the strength of the evidence for a hypothesis given the data, any interpretations herein will be grounded in the classification system established by Jeffreys (1961) and revised by Wasserman (2000). That is, a Bayes Factor between one and three constitutes 'weak' evidence, a Bayes Factor between three and ten constitutes 'moderate' evidence, and a Bayes Factor greater than ten constitutes 'strong' evidence.

Simple effects analyses were conducted by means of Bayesian paired samples t-tests to assess the evidence of blocking within each group. Given that the alternative hypothesis posits a specific direction of the effect (i.e. the presence of blocking: higher scores/ratings for overshadowed than blocked cues) these tests employed an informed alternative hypothesis (see Wagenmakers et al., 2018). Each test specified a one-sided alternative hypothesis assigned a Cauchy distribution with default scaling ($r$ = .707), truncated to allow only positive values of the effect size $\delta > 0$. This was compared against a null hypothesis of no difference between scores or ratings, or $\delta = 0$. All analyses were conducted using the JASP software (JASP Team, 2020; Rouder et al., 2012, 2016, 2017).

Note that since frequentist-based null hypothesis significance testing is more conventional in this area of research, I also include p values calculated from equivalent ANOVAs and t-tests wherever possible. These are provided to facilitate comparison with past research findings, which have been mainly judged according to standard thresholds for statistical significance ($p < .05$). In this research, I will assess relative evidence for differences between conditions versus absence of those differences using the reported Bayes Factors. However, in cases where a hypothesis-relevant result is significant by conventional NHST standards but the evidence from the Bayesian analysis is 'weak' (a Bayes Factor between one and three), I will interpret the result as providing some evidence but will note that it should be treated with caution.

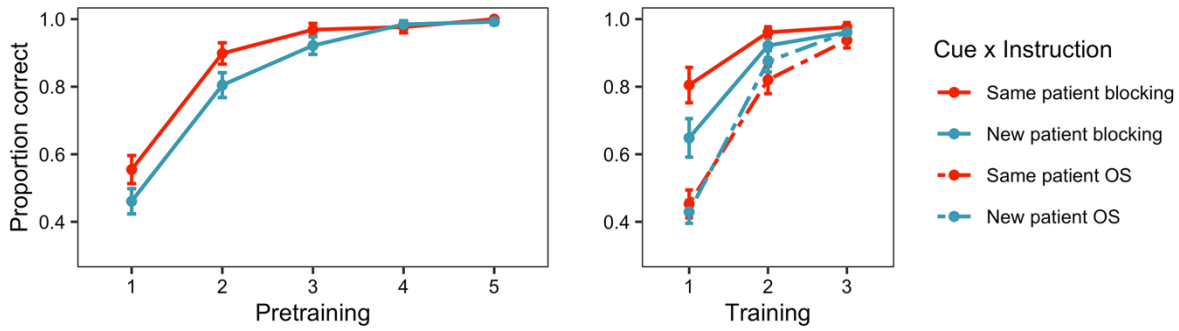
**Results**

Figure 2.1 shows the proportion correct (i.e. accuracy) of predictions made within each block of pretraining and training for the blocking and overshadowing compounds. Accuracy increased steadily across blocks and accuracy in the final block of each phase was greater than 88% for each compound, indicating that all participants learned the contingencies well by the end of training. To assess the effect of the critical instructions on the predictions made during training, I conducted a compound type (blocking vs. overshadowing) by instruction (same patient vs. new patient) Bayesian ANOVA on the accuracy data from the first block of the training phase, immediately after the critical instructions were given. As expected, accuracy was higher for the blocking compounds which contained a pretrained cue than the novel overshadowing compounds on average, $BF_{Incl}$ = 6.706 x $10^6$, $\eta_p^2$ = 0.392, $F(1,62) = 39.95$, $p < .001$. Although numerically the change in patient appeared to reduce the effect of pretraining on prediction accuracy in training, it was not clear from the Bayesian analysis whether the instructions modulate the effect of compound type, $BF_{Excl}$ = 1.453, $\eta_p^2$ = 0.034, $F(1,62) = 2.167$, $p = .146$ in favour of excluding the interaction from the model.[6]

[6] As the overshadowing compounds are completely novel in the training phase, the comparison with the blocking compounds may not be sufficiently sensitive to reveal an effect of instructions that weaken the relevance of pretraining. I therefore repeated this analysis comparing the training predictions for the blocking to the unovershadowing compounds, a complementary cue competition effect in which the pretrained cue predicts a different outcome (no allergic reaction) to the training compound (see Table 2.1 for the contingencies). This showed that the critical instructions appeared to affect training predictions for the blocking and unovershadowing compounds during the initial block of the training phase. A compound (blocking vs. unovershadowing) by instruction (same patient vs. new patient) Bayesian ANOVA revealed that the data were around 2.4 times more likely under a model that included an interaction between compound and instruction, $BF_{Incl}$ = 2.418, $\eta_p^2$ = 0.068, $F(1,62) = 4.49$, $p = .038$. Accuracy was higher for the blocking than the unovershadowing compounds on average, $BF_{10}$ = 5.933 x$10^8$, *error%* = 2.453, $\eta_p^2$ = 0.442, $F(1,62) = 49.12$, $p < .001$. However, the evidence for an interaction indicates that this benefit was less pronounced following new patient instructions. There was no main effect of instruction, $BF_{01}$ = 3.388 in favour of the null, *error%* = 1.106, $\eta_p^2$ = 0.022, $F(1,62) = 1.37$, $p = .247$.

**Figure 2.1**

*Accuracy of pretraining and training predictions in Experiment 2.1.*



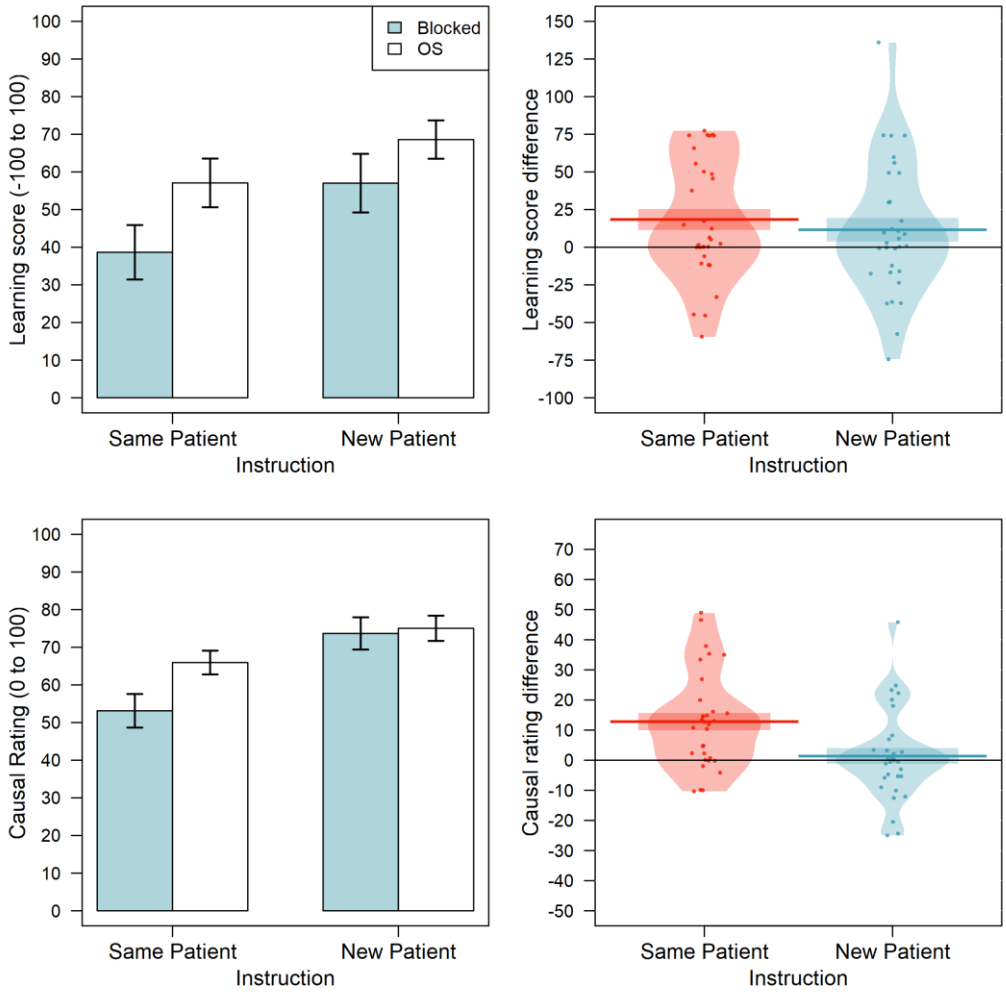

*Note*. Mean proportion correct predictions across both phases of training for the blocking (solid) and overshadowing (OS, dashed), compounds. The same patient group is shown in red, the new patient group in blue. Error bars represent the standard error of the mean.

**Test Phase**

The learning scores and causal ratings for the blocked and overshadowed cues are shown in Figure 2.2. Means for all other cues are reported in Appendix C. Learning scores were subjected to a Bayesian ANOVA with factors of cue (blocking v overshadowing) and instruction (same patient v new patient). The model comparison revealed evidence for a blocking effect (higher learning scores for the overshadowed than the blocked cues), $BF_{10}$ = 7.094, *error%* = 1.640, $\eta_p^2$ = 0.118, $F(1,62) = 8.25$, $p$ = .005. The data did not support a cue by instruction interaction, $BF_{Excl}$ = 3.289, $\eta_p^2$ = 0.007, $F < 1$, indicating that the patient change introduced by instructions did not affect the magnitude of blocking in memory. There was little evidence in either direction when comparing the instruction main effect model against the null model, $BF_{10}$ = 1.124, *error%* = 1.209, $\eta_p^2$ = 0.054, $F(1,62) = 3.53$, $p$ = .065.

**Figure 2.2**

*Learning scores and causal ratings for Experiment 2.1.*




*Note.* Learning scores (top panels) and causal ratings (bottom panels) for the critical test cues as a function of instruction group (same patient vs. new patient) represented in two ways. The left panels illustrate the means (+/- SEM) for the blocked and overshadowed (OS) cues. The right panels show the distribution of the mean difference scores for the blocking effect (overshadowed – blocked). The box and central line represent the standard error of the mean and the mean respectively.

A different pattern emerged when examining the blocking effect in causal ratings. A Bayesian ANOVA revealed evidence in favour of both the main effect model of cue (blocking vs.

overshadowing) and of instruction (same patient vs. new patient) over the null model, $BF_{10}$ = 28.132, *error%* = 0.721, $\eta_p^2$ = 0.180, $F(1,62) = 13.57$, $p < .001$, and $BF_{10}$ = 8.184, *error%* = 1.229, $\eta_p^2$ = 0.120, $F(1,62) = 8.46$, $p = .005$ respectively. The Bayes Factor Inclusion across matched models indicated that the data were 8.981 times more likely under an interaction model than a null model, $\eta_p^2$ = 0.125, $F(1,62) = 8.86$, $p = .004$. This indicates that the magnitude of the blocking effect in causal ratings differed according to which instructions the participants received. Simple effects analyses supported this pattern, with evidence of a blocking effect in the same patient group, $BF_{10}$ = 646, *error%* < 0.001, $d = 0.805$, $t(31) = 4.55$, $p < .001$, whereas in the new patient group the evidence favoured the null, $BF_{01}$ = 3.386, *error%* < 0.001, $d = 0.092$, $t(31) = 0.52$, $p = .304$.

**Discussion**

Consistent with many prior studies investigating blocking in human learning, a reliable blocking effect in causal judgements was observed. However, critically these judgements were sensitive to instructions that manipulated the relevance of pretraining. Specifically, blocking was weakened by instructions that introduced the training phase as a new patient. Only those who received instructions describing the pretraining and training phases as part of the same patient context showed evidence of blocking in their causal judgements. It appears then that instructions can disrupt the expression of blocking in causal judgements. There was some evidence that the instructions changed the way participants made predictions at the start of the training phase. However, since the difference between the blocking and overshadowing compounds was not significant in this experiment, I will reserve discussion of this until later in the chapter.

Considerable evidence exists that people can make logical or reasoned judgements about causality that do not reflect associative memory. However, as others have noted, this does not discount the possibility that associative processes can form the basis of causal judgements under certain circumstances (Shanks, 2007; Le Pelley et al., 2017, Thorwart & Livesey, 2016). What is surprising is that the learning scores did not reflect the same sensitivity to instructions. Associative memory was weaker for the blocked cue than the novel overshadowed control cues, even when the

training phase referred to a new patient. Contrary to the proposal of Thorwart and Livesey (2016), the expectancies important for prediction error learning are seemingly not easily shifted by providing information that calls into question the relevance of associative history. This is even more surprising when we consider that causal judgements (and to some extent overt predictions) do show evidence of critical reflection. This suggests that the operations of the prediction error learning mechanism are somewhat protected from the influence of instructions that can modify causal judgements and to some extent overt predictions. Given that this investigation and pattern of results is novel, it clearly needs replication. Experiment 2.2 sought to provide this, and also to test overt attention during the learning task using measurements of eye gaze.

**Experiment 2.2**

Experiment 2.1 investigated the possibility that instructions affect learning by altering the predictions on which *prediction error* is based. The assumption inherent in the discussion thus far is that prediction error determines the extent to which the *outcome* is processed. That is, in the typical blocking procedure, learning about the blocked cue is poor because there is little prediction error on AB+ trials (in which the blocked cue is presented in compound with a cue that has already been established as a reliable predictor of the outcome). This leads to a failure to process the outcome in a manner that supports learning the blocked cue-outcome association. However, as discussed in Chapter 1, models of attention in learning suggest an alternative mechanism, albeit one which also relies on prediction error. Such models propose that blocking occurs because we fail to process the *blocked cue* on AB+ trials. That is, blocking is the result of learned changes in attention paid to cues. While these two explanations of blocking are not mutually exclusive, proponents of the latter propose that blocking occurs when attention to the blocked cue is weak either because attention is directed to the previously predictive cue A (Mackintosh, 1975) or because the outcome is unsurprising (Pearce & Hall, 1980). There is empirical support for reduced processing of the blocked cue in a blocking procedure. For example, patterns of overt gaze are consistent with poor attention to the blocked cue during training (Beesley & Le Pelley, 2011; Kruschke et al., 2005; Wills et al.,

2007), and blocking appears to be associated with reduced associability, gaze (Beesley & Le Pelley, 2011, Kruschke & Blair, 2000; Le Pelley et al., 2007), and attentional capture (Luque et al., 2018) in subsequent tasks.

The results of Experiment 2.1 suggest that learning scores for specific cue-outcome relationships may not be strongly influenced by instructions. Nevertheless, to the extent that associative memory *is* influenced by instruction, one way in which this influence may be exerted is via modifying the representation of cues via selective attention. For instance, Thorwart and Livesey (2016) argued that non-associative knowledge could impact upon associative learning by bringing selective attention under conscious control and thereby changing the inputs to an associative learning system without modifying the way in which it operates. That is, if instructions can modify selective attention, and if selective attention contributes to blocking in causal learning tasks, then instructions that effectively modify overt attention to cues may affect learning by changing how stimuli that enter into a prediction error learning system are represented.

There is evidence that instructions *similar* to those used in this chapter can influence the allocation of attention and consequently influence learning. For instance, Mitchell et al. (2012) found that the learned predictiveness effect can be reversed with explicit instructions that change participants' expectations about which cues will be predictive in the new context. While the authors proposed that this reversal indicated that reasoning, not associative learning, was driving the learned predictiveness effect, an alternative interpretation is that the instructions led to conscious control of selective attention, which thereby altered later associative learning (Thorwart & Livesey, 2016). Consistent with this interpretation, three subsequent papers found evidence of a reduction in the learned predictiveness effect with such instructions but failed to replicate the complete reversal observed by Mitchell and colleagues (Cobos et al., 2018; Shone, et al., 2015; Don & Livesey, 2015). Most notably, Cobos and colleagues (2018) observed poor control of selective attention. That is, the rapid attentional biases produced by previous learning also persisted under reversal instructions. This lends support to the notion that instruction is a powerful but perhaps not wholly effective tool

when it comes to controlling attention to cues in such tasks. With this in mind, it is possible that the extent to which instructions do or do not influence the strength of associative memory reflects the capacity of those instructions to influence selective attention.

The aim of Experiment 2.2 was to replicate Experiment 2.1 and to investigate whether patterns of selective attention to cues follow the same sensitivity to instructions (and lack thereof). To this end, Experiment 2.2 repeated the core design of Experiment 2.1 whilst also recording eye gaze. For training trials, eye gaze was divided into two periods, the decision period and the feedback period. The decision period was defined as the time between the onset of the cues and the selection of an outcome. Previous studies of attention in cue competition have primarily focussed on this window as the period in which attentional shifts are relevant, that is, when the participant reencounters the cue and forms a prediction about the outcome. This is true (albeit implicitly) of many attentional models (Mackintosh, 1975; and Pearce & Hall 1980, for example) but the EXIT model of attention proposes that prediction error learning produces immediate attentional shifts after a prediction is made, as soon as prediction error is encountered (Kruschke, 2001). These shifts are proposed to reflect updating following an error, and are directed towards cues that will minimise error on that trial. Don and colleagues (2019) detected attentional shifts in the feedback period consistent with the predictions of the EXIT model in the inverse base rate effect (a decision bias to which attention-based prediction error models have been successfully applied, see Don et al., 2021 for a review). For consistency, an analysis of eye gaze in the feedback period is included in Appendix D but the analyses reported here focus on the prediction period.

In light of the evidence outlined above, I expected the memory deficits for the blocked cue observed in the previous experiments to be accompanied by a bias in attention toward the pretrained cue A on AB trials, which is established as a good predictor in pretraining. These predictions should hold in the same patient condition. However, if explicit instructions effectively modify selective attention, then any biases in overt attention on blocking trials should reflect the influence of the change in patient. That is, the new patient instructions should weaken the

attentional biases that would otherwise be present in training. This would be consistent with Mitchell et al.'s (2012) findings. Overt attention to cues should match decisions made on a given trial, thereby reflecting sensitivity to the instructed context change in the same way that overt predictions do. Nonetheless, the evidence from Experiment 2.1 suggests that this may not be the case.

## Method

### Participants

Sixty-four undergraduate psychology students from the University of Sydney participated in the study in exchange for partial course credit. As the modifications to the design increased the difficulty of the task, the learning criterion was relaxed to 50% correct responses across all pretraining trials (chance performance = 25% correct). Following this, two participants were excluded from analysis, both from the New Patient condition. The final sample of 62 (55 female, *M* age = 21.33), were randomly assigned to either the same Patient (*n* = 32) or new Patient (*n* = 30) condition.

### Design Apparatus & Stimuli

The design followed that of Experiment 2.1, however in order to increase the opportunities to derive reliable eye gaze data the number of iterations of the critical contingencies was increased from two to three. This change necessitated adding a third outcome, "Headache," and four additional food cues. The same 300 x 300 pixel images served as the food cues however, to create a clear delineation between the cue and the background, a black border was added to the cues and the text labels were not included. In addition to the individual cue test, a separate test phase was added with negation trials that pitted two critical cues that were paired with different outcomes during training against one another. The design of Experiment 2.2 is depicted in Table 2.2.

**Table 2.2**

*Design of Experiment 2.2*

| | | | Test | |
|---|---|---|---|---|
| | Pretraining | Training | Individual | Negation |
| Blocking | A – O1 | AB – O1 | A, B | BI |
| | C – O2 | CD – O2 | C, D | DK |
| | E – O3 | EF – O3 | E, F | FG |
| Overshadowing | | GH – O1 | G, H | PB |
| | | IJ – O2 | I, J | RD |
| | | KL – O3 | K, L | NF |
| Unovershadowing | M – no O | MN – O1 | M, N | LN |
| | O – no O | OP – O2 | O, P | PH |
| | Q – no O | QR – O3 | Q, R | JR |
| Fillers | ST – O1 | Y – noO | Y | |
| | UV – O2 | Z – no O | Z | |
| | WX – O3 | | | |
| | a – no O | ab – no O | a, b | |
| | | cd – no O | c, d | |

*Note.* Letters A-d represent the cues randomly assigned to different food items for each participant. Outcomes O1-O3 were randomly assigned to *headache, nausea,* and *fever*; 'no O' denotes *no reaction.*

Participants were tested individually using a Tobii TX300 eye tracker mounted on a 23inch widescreen monitor (Tobii technology, Sweden). Gaze location was recorded at a sampling rate of 300 Hz while participants completed the allergist task. Participants were positioned approximately 65cm from the monitor with a desk-mounted chinrest.

**Procedure**

The eye tracker was calibrated for each participant at the beginning of the experiment with a five-point procedure. Each trial commenced with a 500ms fixation cross that appeared in the

horizontal centre of the screen, at the vertical position of the cues. Training otherwise proceeded as described in Experiment 1. Participants completed five blocks of pretraining and three blocks of training, with each contingency presented twice per block. The procedure for the recall and causal ratings test was identical to the test phase of Experiment 2.1, with the exception that the recall test included the additional outcome, such that participants were asked to make a rating for each of the *four* possible outcomes from training (*fever*, *nausea*, *headache*, and *no reaction*).

Following the individual cue test phase, participants completed a negation test in which they were presented with foods from the training phase in the novel combinations shown in Table 2.2. On each trial, two cues were presented, and participants were asked to select the outcome that the patient from training was most likely to experience as a result of eating these new meals. The left-right configuration of the cues was counterbalanced across participants. After making a selection, participants rated their confidence on a ten-point linear analogue scale ranging from "not at all confident" to "very confident." Participants were required to confirm their selections by way of a key press in order to proceed to the next trial, this enabled them to adjust their initial responses if they desired.

**Eye gaze analysis**

The eye gaze analyses focused on the decision period between the stimulus onset and the response for training and negation trials. Gaze was defined as falling on a cue if it fell within a boundary 100 pixels outside the border of the cue image (500 x 500 pixels). The tracker recorded gaze location for each eye separately. I calculated the proportion of missing data for each eye on each trial and retained only the eye with the least missing data for that trial. For gaps of less than 75 milliseconds, missing data from the retained eye was interpolated from the immediately surrounding data. A displacement method was used to determine fixations separately for the two periods of interest (Salvucci & Goldberg, 2000). Gaze was analysed in windows of 150ms with a displacement criterion of 75 pixels. That is, a window was deemed a fixation if the X and Y coordinates within that window did not deviate beyond a range of 75 pixels. Fixation duration was

determined by extending the fixation window until there was a displacement greater than 75 pixels. The position of the fixation was given by the mean of the X and Y coordinates across the fixation sample.

Trials with no recorded fixations for the relevant period were removed from the analyses (*M* = 2.29%, *SE* = 0.61 for the decision period of training, *M* = 3.18%, *SE* = 0.53 for the feedback period of training, and *M* = 2.08 %, *SE* = 0.53 for negation trials). Any participants with more than 30% of trials missing eye gaze data for the phase under investigation were excluded from further analysis. For this reason, one participant was excluded from the analysis for the training phase and one participant from the negation phase. For the remaining participants, any trials with no recorded gaze on either cue were also excluded from the analysis (*M* = 12.76%, *SE* = 2.02 for the decision period of training, *M* = 22.71%, *SE* = 1.85 for the feedback period of training, and *M* = 7.54%, *SE* = 1.79 of negation trials). To control for differences in response latencies across trials, dwell time (summed duration of fixations on each cue) was analysed as a proportion of response time for the decision period. These proportions were aggregated over cues of the same type, for instance proportion dwell time on the blocked cue is the mean of the proportion dwell time for cues B, D and F.

Consistent with the analyses for the overt predictions made during training, I conducted a focused analysis on gaze patterns in the initial block of the training phase. For this analysis I calculated a bias score, which captures the dwell time on the target (the blocked cue) as a proportion of the summed dwell time on both the target and pretrained cues (we will call this the proportion target score, Beesley & Le Pelley, 2011; Wills et al., 2007). A value of 0.5 indicates no bias, whereas values greater than or less than 0.5 indicate a bias toward, or away from the target cue respectively.

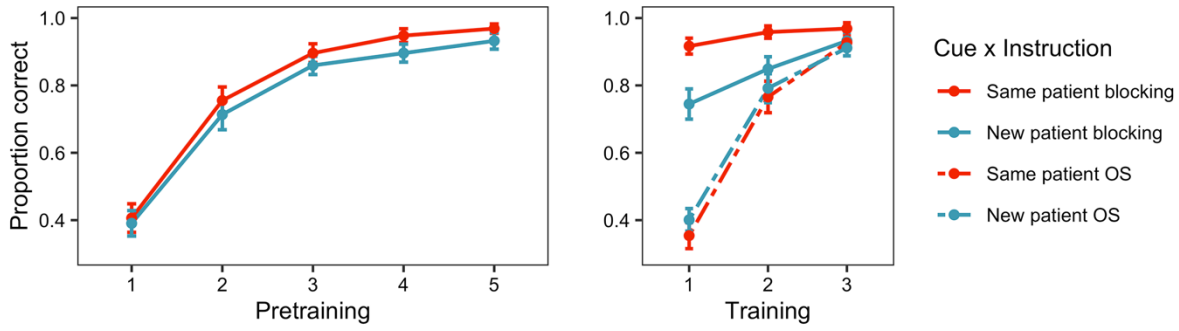
## Results

### Training

Figure 2.3 shows the mean proportion correct outcome predictions averaged over equivalent types of trials, during each training stage. There is clear evidence of learning in both phases with accuracy of predictions reaching above 85% for all trial types by the end of each training phase. Comparing performance in the block immediately following the critical instructions, there was evidence that the instruction affected the extent to which pretraining predictions were carried forward into training. Accuracy was on average higher for the blocking compounds than the overshadowing compounds (which were novel at the beginning of training), $BF_{10}$ = 9.328 x $10^{23}$, *error%* = 1.203, $\eta_p^2$ = 0.760, $F(1,60)$ = 190.48, $p$ < .001. However, as illustrated in Figure 2.3, the benefit for blocking compounds was less pronounced when a change in patient was introduced, $BF_{Incl}$ = 25.350, $\eta_p^2$ = 0.147, $F(1,60)$ = 10.330, $p$ = .002. The evidence did not support a main effect of instruction, $BF_{10}$ = 0.301, *error%* = 1.352, $\eta_p^2$ = 0.031, $F(1,60)$ = 1.913, $p$ = .172.

**Figure 2.3**

*Accuracy of pretraining and training predictions in Experiment 2.2*



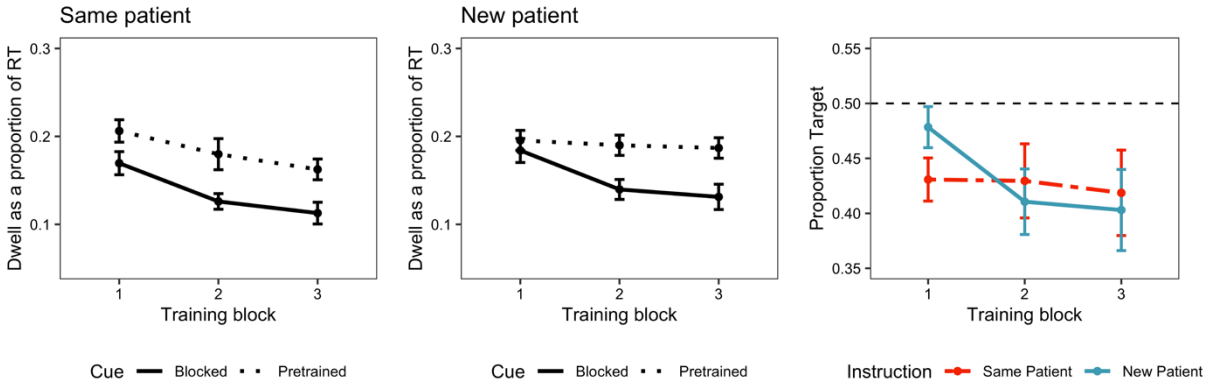

*Note.* Mean (+/- SEM) proportion correction predictions during for the blocking (solid lines) and overshadowing ("OS", dashed lines) compounds. Shown as a function of instruction group (same patient in red, new patient in blue).

***Eye gaze***

Figure 2.4 shows mean dwell times for the decision period as a proportion of response time for the blocking compounds across the training phase. To determine whether there was a bias in attention toward the pretrained cues (away from the blocked cues), the proportion dwell time for the pretrained (A, C, and E) and blocked (B, D, and F) cues of the blocking compounds were subjected to a Bayesian ANOVA with factors of block (1-3), cue (pretrained vs. blocked), and instruction (same patient vs. new patient). This yielded strong evidence for a main effect of cue, $BF_{10}$ = 4.753 x$10^8$, *error%* = 10.084, $\eta_p^2$ = 0.351, $F(1,57) = 30.87$, $p < .001$, indicating that overall participants directed greater overt attention to the pretrained cues than the blocked cues. This bias did not appear to vary across blocks ($BF_{Excl}$ = 2.150, $\eta_p^2$ = 0.038, $F(2,114) = 2.25$, $p = .110$) nor instruction groups ($BF_{Excl}$ = 4.814, $\eta_p^2$ = 0.006, $F < 1$). There was moderate evidence against a three-way interaction, $BF_{Excl}$ = 6.635, $\eta_p^2$ = .009, $F < 1$. Strong evidence for a linear effect of block with no block by instruction interaction suggests that attention decreased across training at the same rate for both groups, $BF_{10}$ = 8360, *error%* = 1.127, $\eta_p^2$ = 0.334, $F(2,114) = 28.54$, $p < .001$, and $BF_{Excl}$ = 6.731, $\eta_p^2$ = 0.032, $F(2,114) = 1.88$, $p = .158$ respectively.

**Figure 2.4**

*Eye-gaze for the decision period of training trials in Experiment 2.2.*




*Note.* The left and centre panels show mean dwell time on blocked and pretrained cues as a proportion of response time for the same patient group (left) and the new patient group (centre). The right panel shows the dwell time for the blocked cues as a proportion of the summed dwell time for the blocked and pretrained cues (proportion target) as a function of instruction group. Error bars represent standard error.

To determine whether there was an early bias in attention away from the blocked cue that differed as a function of the instructions that participants received, I examined the proportion target scores in the first block of the training phase, immediately following the critical instruction. Proportion target scores for each instruction group are shown in the third panel of Figure 2.4. Although the bias was numerically weaker in the new patient group, a Bayesian ANOVA indicated that the evidence for a difference in the size of the bias was equivocal, $BF_{10}$ = 0.941, *error%* = 0.009, $\eta^2$ = 0.050, $F(1,59) = 3.084$, $p = .084$. Nonetheless, comparing the proportion target scores in each instruction group against a value of 0.5 (reflecting no gaze bias or dwell time split equally across cues on a trial), there was a bias in the direction of the pretrained cue in the same patient group, $BF_{10}$ = 25.032, *error%* < .001, $d = 0.634$, $t(30) = 3.531$, $p < .001$, but the evidence weakly favoured the null hypothesis that there was no bias in gaze towards the pretrained cue in the new patient group, $BF_{01}$ = 2.811, *error%* = .004, $d = 0.211$, $t(29) = 1.154$, $p = .129$.

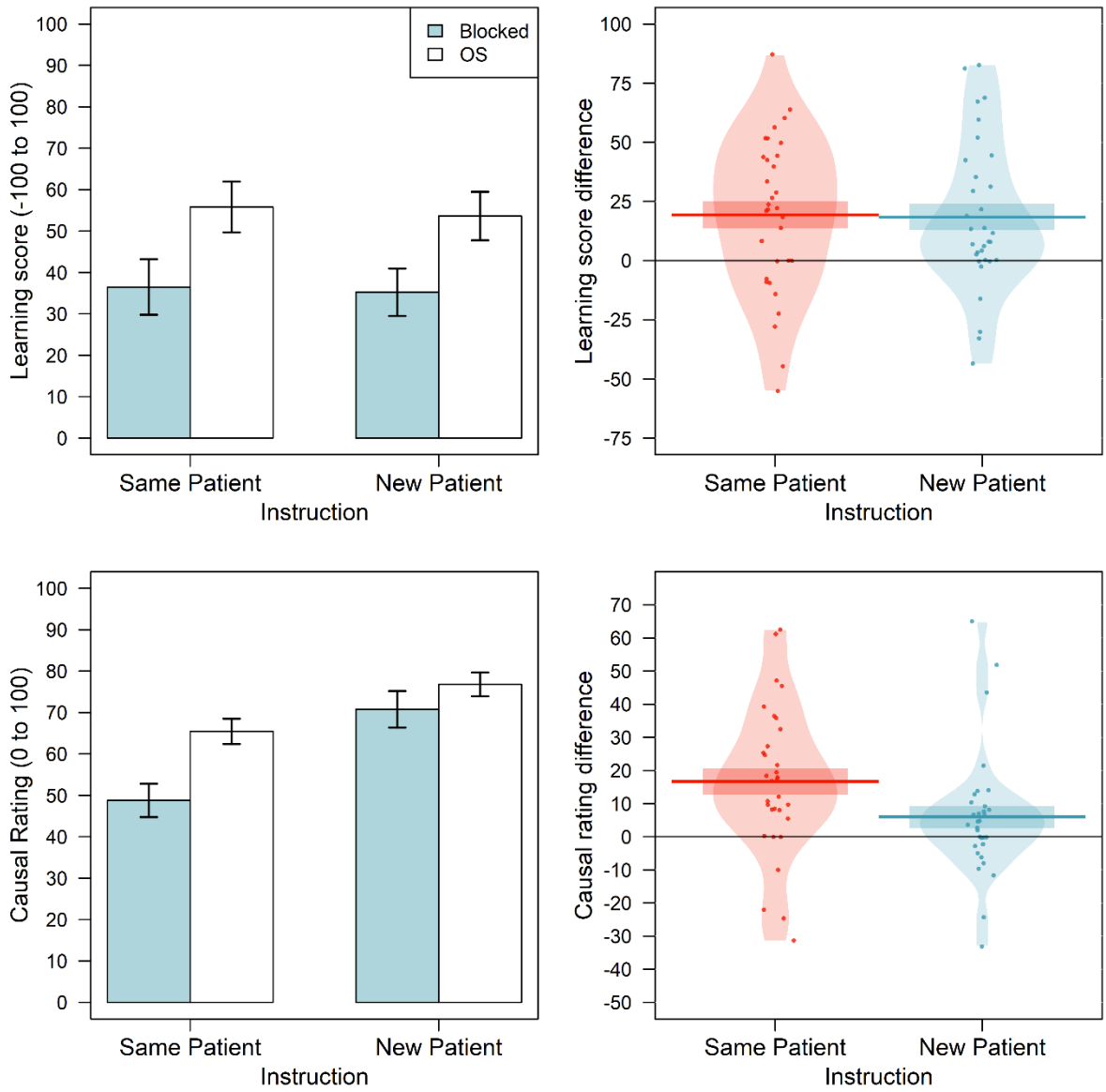
**Test Phase**

Figure 2.5 depicts the learning scores and causal ratings for blocking and overshadowing compounds from the test phase of Experiment 2.2. Consistent with Experiment 2.1 there was a blocking effect in learning scores, $BF_{10}$ = 1038, *error%* = 1.315, $\eta_p^2$ = 0.263, $F(1,60) = 21.436$, $p < .001$, that did not appear to interact with instruction, $BF_{Excl}$ = 3.861 across matched models, $\eta_p^2 < .001$, $F < 1$. The data favoured the null over the main effect of instruction, $BF_{01}$ = 3.471, *error%* = 1.811, $\eta_p^2 < .001$, $F < 1$.

In causal ratings there was again strong evidence for a blocking effect, $BF_{10}$ = 371, *error%* = 1.718, $\eta_p^2$ = 0.243, $F(1,60) = 19.257$, $p < .001$. However there was also a main effect of instruction, $BF_{10}$ = 64.113, *error%* = 1.461, $\eta_p^2$ = 0.188, $F(1,60) = 13.908$, $p < .001$ indicating that the causal ratings were on average higher for those in the new patient than the same patient group. The evidence for retaining an interaction between cue and instruction was equivocal, $BF_{Incl}$ = 1.114, $\eta_p^2$ = 0.057, $F(1,60) = 3.367$, $p = .061$. Follow up tests revealed that the same patient group showed clear evidence of a blocking effect in causal ratings, $BF_{10}$ = 310, *error%* < 0.001, $d = 0.754$, $t(31) = 4.264$, $p < .001$. For the new patient group, however, the evidence did not strongly favour either hypothesis, $BF_{01}$ = 1.692 toward the null, *error%* = 0.005, $d = 0.339$, $t(29) = 1.856$, $p = .037$, indicating weaker evidence of blocking than for the same patient group.

**Figure 2.5**

*Learning scores and causal ratings for Experiment 2.2*




*Note.* Learning scores (top panels) and causal ratings (bottom panels) for the critical test cues as a function of instruction group (same patient vs. new patient) represented two ways. The left panels illustrate the means (+/- SEM) for the blocked and overshadowed (OS) cues. The right panels show the distributions of the mean difference scores for the blocking effect (overshadowed – blocked). The central bold line and surrounding box represent the mean and standard error of the mean respectively.

***Negation scores***

Negation scores for the critical comparison, blocked vs. overshadowed, were calculated by taking the proportion of trials on which participants selected the relevant outcome for each element of the negation compound. Table 2.3 shows the proportion choice for the outcome paired with the blocked cue, the overshadowed cue, and no allergic reaction. I conducted a Bayesian repeated measures ANOVA comparing choice proportions for the relevant outcomes across instruction groups. There was a choice bias towards the outcome paired with the overshadowed cue, $BF_{10}$ = 3.296 x$10^{11}$, *error%* = 1.141, $\eta_p^2$ = 0.414, $F(1,60) = 42.45$, $p < .001$. Crucially the magnitude of this bias appeared to be equivalent across instruction groups, $BF_{Excl}$ = 3.830, $\eta_p^2$ = 0.001, $F < 1$ against the interaction. In other words, the pattern of responding on the negation trials was similar to that of the learning scores.

**Table 2.3**

*Proportion choice and confidence ratings for negation trials*

| | Choice | | | |
|---|---|---|---|---|
| Group | Blocked | OS | No reaction | Confidence |
| Same patient | 0.208 | 0.635 | 0.083 | 62.99 |
| New Patient | 0.233 | 0.689 | 0.022 | 64.48 |

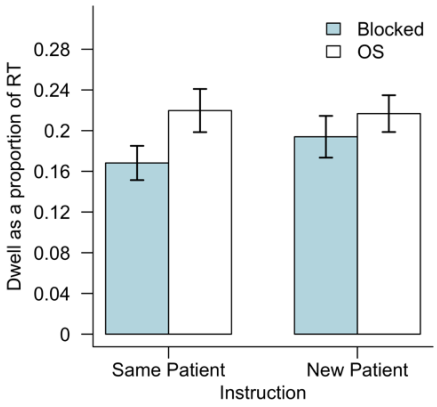
*Note*. This table shows the data from the negation trials that pitted a blocked cue against an overshadowed cue (BI, DK, FG) as a function of instruction group (same patient vs. new patient). Choice columns show the proportion choice of the outcome paired with the blocked cue, overshadowed cue (OS), or the "no reaction" outcome along with the mean choice confidence. Proportions do not sum to one as there was a fourth choice not included in this table, the allergic reaction that was not paired with either cue.

***Negation eye gaze***

Mean dwell times as a proportion of reaction time on the negation trials are shown in Figure 2.6. Consistent with the behavioural data, numerically there appeared to be a gaze bias away from the blocked cue indicating participant spent more time looking at the overshadowed cue on these trials. However, the evidence for this bias was equivocal, $BF_{10}$ = 1.614, *error%* = 0.923, $\eta_p^2$ = 0.067, $F(1, 62) = 4.431$, $p = .039$. This did not differ as function of instructions, $BF_{Excl}$ = 3.024, $\eta_p^2$ = 0.001, $F < 1$, nor did the instructions affect overall dwell time, $BF_{Excl}$ = 3.933, $\eta_p^2$ = 0.005, $F < 1$.

**Figure 2.6**

*Eye-gaze for the decision period of negation trials in Experiment 2.2.*




*Note.* Bars represent the mean (+/- SEM) dwell time as a proportion of reaction time. Means are depicted as a function of instruction group (same patient vs. new patient) for trials pitting blocked cues against overshadowed (OS) cues.

## Discussion

Experiment 2.2 replicated the key findings of Experiment 2.1. The results provided stronger evidence that the active predictions made during training were subject to instructed control, such that introducing a new patient weakened the effect of pretraining on overt predictions in training. The same instructions also modified the strength of the blocking effect in causal judgements made at test, although the evidence for this was substantially weaker here than in Experiment 2.1. Finally,

associative memory judgements were again not affected by the instructed relevance of pretraining. That is, while overt predictions made during training, and to some extent causal judgements, appeared to reflect the influence of explicit instructions, there was nonetheless a robust blocking effect in associative retrieval at test that was insensitive to the instructed context change. This was complemented by a persistent blocking effect in choice behaviour on negation trials regardless of the instructions given, suggesting that the overshadowed cues had greater control over outcome choices than the blocked cues.

In Experiment 2.2 I also recorded eye gaze on the cues during the training and negation phases. Consistent with past research, in the training phase attention was biased towards the pretrained cue for the blocking compounds. It was not clear, however, whether this bias was sensitive to the instructed relevance of past learning. On the one hand, an analysis of the trends in gaze during training indicated there was a bias in gaze toward the pretrained cue in both groups that remained stable throughout suggesting that the instructions had little effect on the attention paid to cues. Similarly, there was a blocking effect in attention to cues on negation trials that directly pitted blocked against overshadowed cues, and this appeared to be robust to the effect of instructions. On the other hand, considering only the first block of training in which the opportunity to detect an effect of instructions should be greatest, there was some evidence that gaze patterns were consistent with the notion that the new patient instructions successfully moderated how participants deployed attention on blocking trials. That is, though the evidence for an interaction with instructions was equivocal, only those who saw the same patient in the training phase showed evidence of a bias consistent with pretraining. The new patient instructions appeared to temporarily eliminate the bias in selective attention in the early stages of training. This pattern is akin to that observed in the overt predictions that participants made during training; there was some early sensitivity to instructions in eye gaze and predictions. I will return to the broader implications of these eye gaze results in the General Discussion.

## Experiments 2.3a and 2.3b

In the previous experiments associative memory was notably resistant to instructed control. There was a pattern of blocking in learning scores that did not interact with instructions about the relevance of pretraining. Notably, however, these same instructions did reduce the effect of pretraining on the magnitude of blocking in participants' judgements of causality. Blocking was weaker in causal ratings when participants were instructed to treat the training phase as a new patient than when the training phase was described as a continuation of pretraining. Thus, although the instructions successfully reduced the perceived relevance of pretraining (evident in causal judgements), they appeared to have had little effect on how the cue-outcome relationships were encoded in memory. However, the inclusion Bayes Factors indicate that the strength of the evidence that instructions affected blocking in causal ratings is substantially weaker in Experiment 2.2 ($BF_{Incl}$ = 8.981 in Experiment 2.1, and $BF_{Incl}$ = 1.114 in Experiment 2.2). This could indicate an issue with sample size in the previous experiments. The decisions about sample size were guided by past research and considerations of feasibility but due to the inconsistency of the results it was necessary to replicate the findings with a larger sample. A compromise power analysis on the interaction in question (taking an average of the observed effect sizes) yielded implied power of 0.99 with a sample size of N = 160. Experiments 2.3a and 2.3b aimed to replicate Experiment 2.1 with this larger sample size.

With the current experiments I aimed to address two additional potential weaknesses in the previous experiments. First, the order of the test questions was fixed, participants always completed the memory ratings before the causal judgements. This could produce an order effect in responses, thereby reducing confidence that the observed dissociation between associative memory and causal judgement is attributable to the impact of the instruction manipulation alone instead of, for example, higher sensitivity to the instruction on the second question because participants have had longer to think about the cues by the time they make this rating. In the following experiments the order of the test questions was counterbalanced. This was taken a step further in Experiment 2.3a

by separating the memory and causal judgements into two blocks and counterbalancing the order in which participants completed the blocks. Second, in the previous experiments the overall base rate of allergic reactions during pretraining was lower than the base rate during the training phase. This could affect the relative strength of blocking in the instruction groups as participants in the same patient group would estimate the base rate across pretraining and training whereas those in the new patient group were encouraged to disregard the pretraining and should therefore only consider the base rate of allergic reactions in the training phase when making a judgment about the likelihood that a cue is allergenic. Given the evidence that causal judgements are sensitive to the base rate of events, particularly for ambiguous cues or those for which participants do not have a strong memory (Jones et al., 2019), we might expect higher causal ratings in the new patient group on average and therefore less evidence of blocking in the new patient group, which is consistent with the results of previous experiments. As a result, Experiment 2.3b matched the base rate of allergic reactions in the pretraining phase to that of the training phase with the addition of some filler cues.

In applying the instruction manipulation, I am making the assumption that when participants are told that pretraining and training stages are two different patients, then what has previously been learned about cue-outcome relationships in the pretraining stage should be considered far less relevant to the cues presented in the subsequent training stage. One could argue that under these instructions, pretraining involving Miss Y's allergies should be *completely* irrelevant to those of Mr X, meaning that it would be quite rational for the blocking effect to completely disappear (and by extension it is possibly *irrational* for it to persist under these conditions). However, it is also defensible to assume that the learner may take the evidence they have gained from observing Miss Y as being relevant—within the confines of this hypothetical scenario—to the overall likelihood that certain foods will lead to certain allergic reactions. Experiment 2.3b included a short survey to test whether participants endorsed these assumptions.

## Method

### Participants

I aimed to recruit 320 participants from the undergraduate psychology subject pool at the University of Sydney to participate in exchange for a small amount of credit towards their assessment. A total of 166 students participated in Experiment 2.3a, and of these, eight were excluded for failing to meet the learning criterion (six of whom had received the new patient instructions). This left a final sample of 158 (M age = 20.3 years, 114 female), randomly assigned to the same patient (n = 82) and new patient (n = 76) instruction groups. Of the 160 participants recruited for Experiment 2.3b, two did not meet the learning criterion and were excluded (one from each instruction group). The remaining 158 participants (*M* age = 20 years, 118 female) were randomly allocated to instruction groups (*n* = 81 same patient, *n* = 77 new patient).

### Design & Procedure

The design and procedure for both experiments was identical to that of Experiment 2.1 with the following exceptions. For Experiment 2.3a the procedure of the test phase was modified such that the two successive test judgements (recall of the outcome and causal judgement) were separated into two test blocks, a memory test block and causal test block. In previous experiments participants were told that they would be making two different kinds of judgements about the cues at the beginning of the test phase and the distinction between these judgements was emphasised in the instructions. To test whether the dissociation between judgements would survive even when these two judgements were not placed in direct contrast, participants were instructed to complete one set of judgements (e.g., the memory test) before they were informed that they would be making a second kind of judgement (e.g. causal ratings) about the cues. The order of the test blocks (memory test first or causal test first) was counterbalanced across participants.

Experiment 2.3b returned to successive test judgements but counterbalanced across participants the order in which they were completed. Two additional filler cues, presented individually and followed by no reaction, were added to the training phase of Experiment 2.3b such

that the probability of the patient suffering an allergic reaction on a given trial was 0.66 in both pretraining and training. Experiment 2.3a did not include this modification.

Finally, Experiment 2.3b included a short survey to test the extent to which participants assume that one patient's allergies are informative about a different patient's allergies. After they completed the test judgements participants were given a brief hypothetical scenario which described a patient ('Patient 1') with a known allergy to grapes, a food not encountered in the previous tasks (the full script can be found in Appendix A). Participants were then asked to rate the extent to which they agreed with three statements about the impact of this information (the grape allergy of Patient 1) on the assessment of the allergies of a novel patient (Patient 2). Responses were made on a 5-point Likert scale ranging from 'strongly disagree' to 'strongly agree.' The survey items are reported below with the results.

## Results

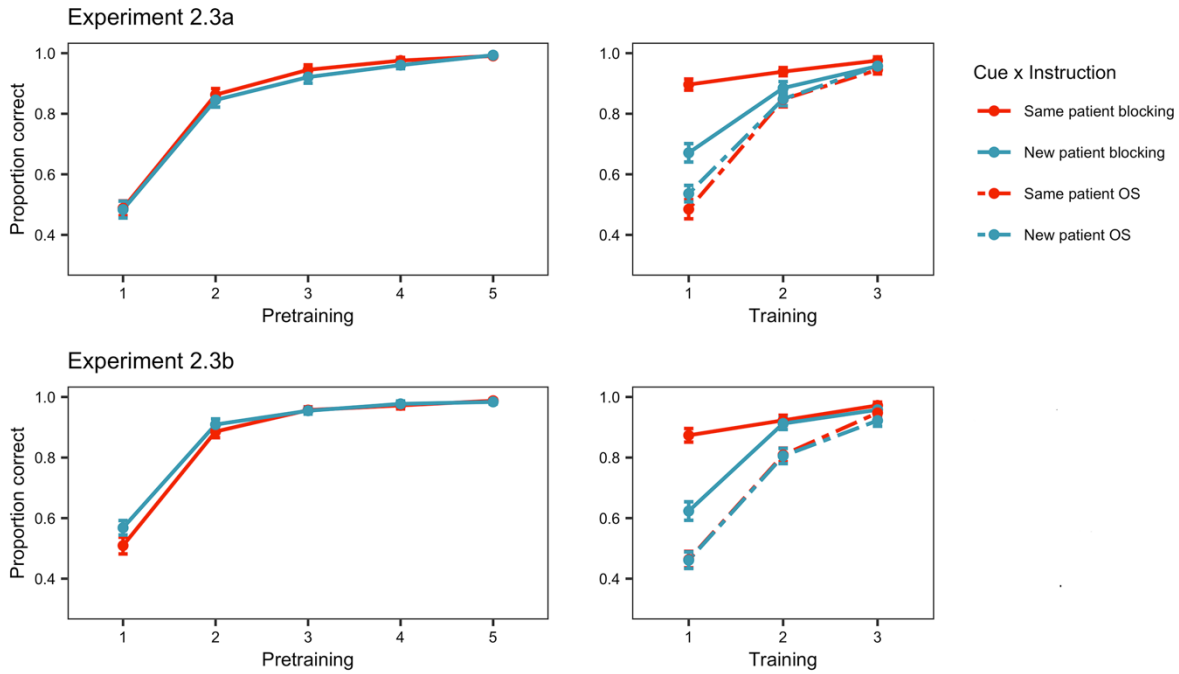
### Training

Accuracy in the final blocks of pretraining and training was high (above 90% on average) indicating sufficient learning of the contingencies in both Experiments 2.3a and b. Figure 2.7 shows the mean accuracy for the blocked and overshadowed cues from Experiment 2.3a (top panels) and 2.3b (bottom panels) as a function of training phase, block, and instruction group. In line with previous experiments, I compared accuracy for the blocked and overshadowed compounds in the initial block of the training phase after participants had received the critical instructions. A compound (blocking vs. overshadowing) by instruction group (same patient vs. new patient) by test order (memory first vs. causal first) by experiment (2.3a vs. 2.3b) Bayesian ANOVA indicated strong support for an effect of compound, $BF_{10} = 1.862 \times 10^{38}$, *error%* = 1.558, $\eta_p^2 = 0.413$, $F(1,308) = 216.65$, $p < .001$, and of instruction group, $BF_{10} = 1324$, *error%* = 1.404, $\eta_p^2 = 0.088$, $F(1,308) = 29.76$, $p < .001$. However, notably there was evidence of an interaction between compound and instruction, reflecting a weaker effect of pretraining when the training patient was described as a new patient, $BF_{Incl} = 4.362 \times 10^9$, $\eta_p^2 = 0.134$, $F(1,308) = 47.55$, $p < .001$. This pattern did not appear to differ as a function of test

order or experiment as the null hypothesis was 5.335 times more likely than the alternative of a four-way interaction, $\eta_p^2 < 0.001$, $F(1,308) = 0.002$, $p = .968$. The remaining main effects and interactions were not supported by the data, smallest $BF_{Excl} = 1.020$ ($\eta_p^2 = 0.013$, $F(1,308) = 4.01$, $p = .046$) for the instruction group x test order x experiment three-way interaction.

**Figure 2.7**

*Accuracy of pretraining and training predictions in Experiments 2.3a and 2.3b*




*Note.* Mean (+/- SEM) proportion correct predictions across training as a function of compound type (blocking vs overshadowing), instruction group (same patient vs new patient) and experiment.

**Test**

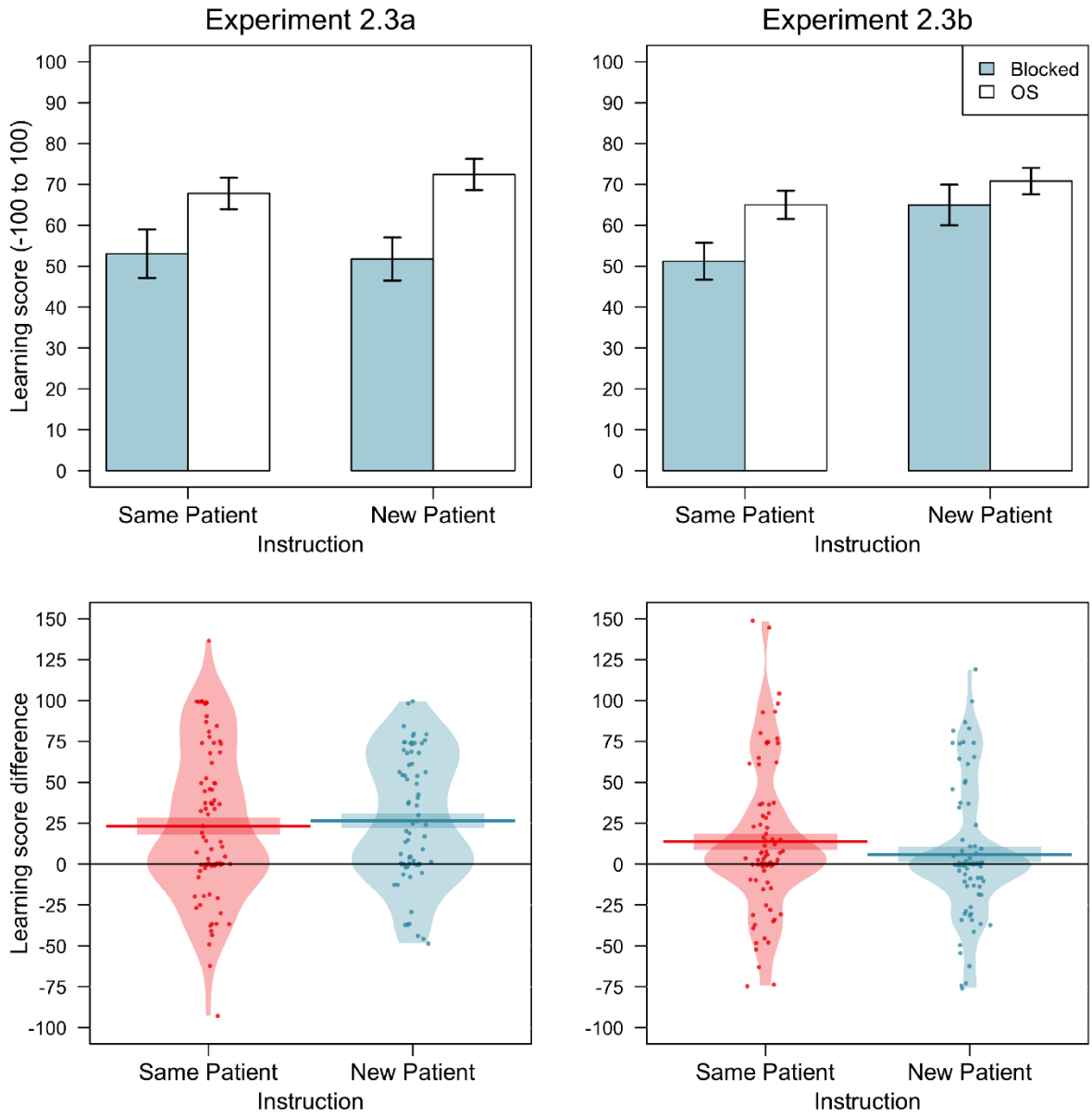
I conducted separate repeated measures Bayesian ANOVA on the learning scores and the causal ratings, each with cue (blocked vs. overshadowed) as a within-subjects factor and instruction group (same patient vs new patient), test order (memory first vs causal first), and experiment (a vs b) as between-subjects factors. The learning scores for the critical test cues in each experiment are given in Figure 2.8. The violin plots show the distribution of the difference scores for the learning

scores and ratings, with difference scores above zero representing a blocking effect (higher scores/ratings for the overshadowed cue than the blocked cue). For learning scores, the ANOVA indicated that the data were 27,801 times more likely under the alternative hypothesis that the learning scores were lower for blocked than overshadowed cues, than under the null hypothesis of no difference (*error%* = 1.287, $\eta_p^2$ = 0.080, $F(1,308) = 26.76$, $p < .001$). The evidence suggested that the magnitude of blocking was not modified by experiment, test order, or instruction group as the Bayes Factors for all interactions favoured the null hypothesis, smallest $BF_{Excl}$ = 1.901 ($\eta_p^2$ = 0.008, $F(1,308) = 2.51$, $p = .114$) for the cue by instruction group by experiment interaction[7]. The data were 5.22 times more likely under a model that excluded the four-way interaction ($\eta_p^2 < 0.001$, $F(1,308) = 0.001$, $p = .973$), and most notably, were consistent with previous evidence that instructions did not modulate blocking in learning scores, $BF_{Excl}$ = 7.961, $\eta_p^2 < 0.001$, $F(1,308) = 0.007$, $p = .934$ for the cue type x instruction group interaction.

[7] The reader may note that the blocking effect is numerically weaker in the new patient group in Exp 2.3b but there is insufficient statistical evidence to support this.

**Figure 2.8**

*Learning scores for Experiments 2.3a and 2.3b*




*Note*. Left panels show learning scores from Experiment 2.3a, right panels show scores from 2.3b. The columns (top row) show the mean (+/- SEM) for the blocked and overshadowed cues. The violin plots (bottom row) show the distributions of blocking scores (mean difference overshadowed – blocked). Individual blocking scores are represented by points. The mean and standard error are given by the horizontal bar and surrounding shaded box respectively. Learning scores are represented as a function of instruction group (same patient vs. new patient).

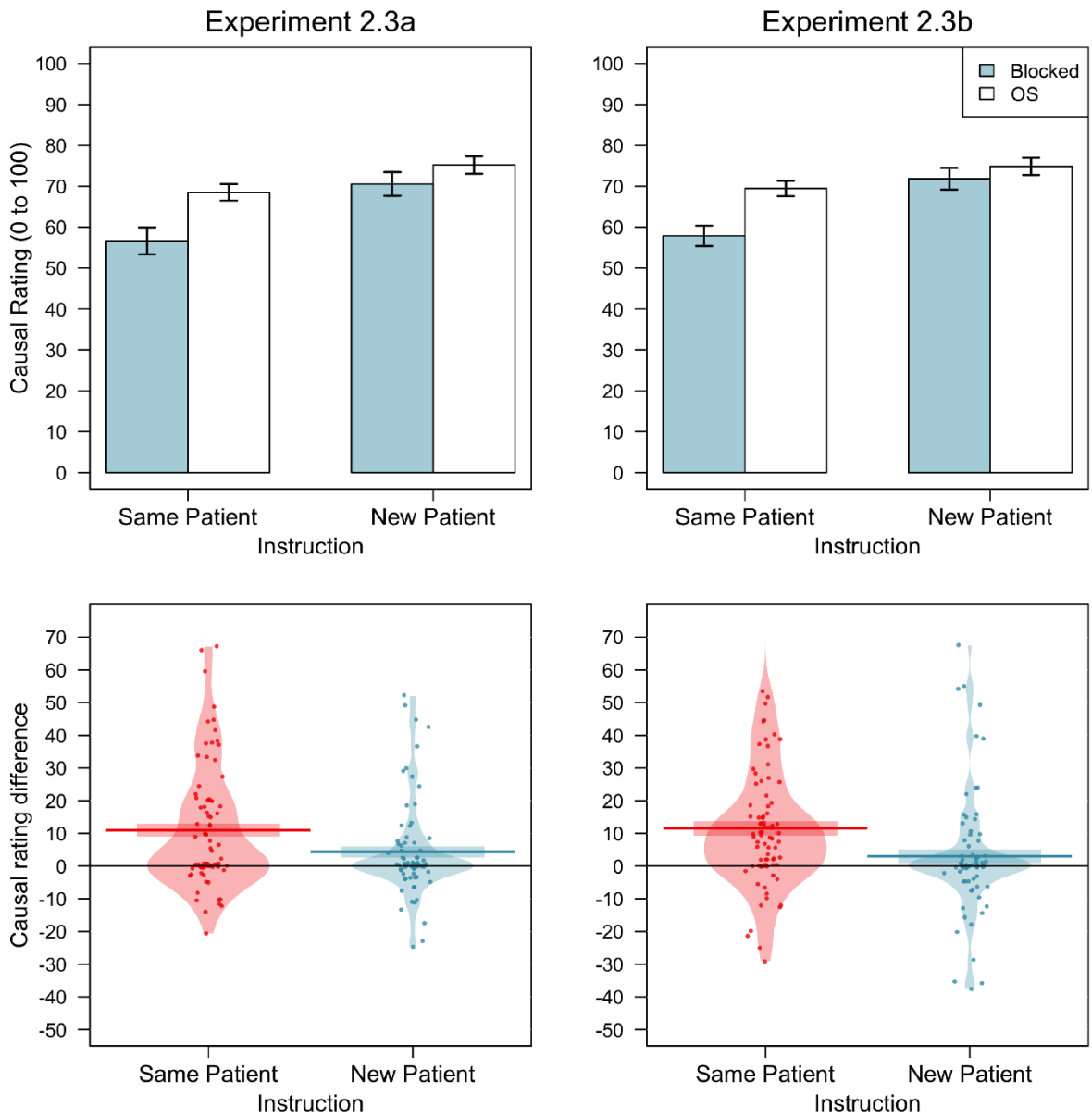
Causal ratings are shown in Figure 2.9. The analysis again yielded strong evidence for a main effect of cue type in the direction of a blocking effect, $BF_{10}$ = 4.128 x $10^{10}$, *error%* = 2.666, $\eta_p^2$ = 0.164, $F(1,308) = 60.41$, $p < .001$. However, notably the evidence supported an interaction between cue type and instruction group, $BF_{incl}$ = 188.98, $\eta_p^2$ = 0.048, $F(1,308) = 15.39$, $p < .001$, supporting the pattern observed in previous experiments of weaker blocking in causal ratings following the new patient instructions than same patient instructions. Follow up Bayesian t-tests within each instruction group indicated the presence of a blocking effect in the same patient instruction group, $BF_{10}$ = 1.492 x $10^{10}$, $d = 0.607$, $t(162) = 7.75$, $p < .001$. However, although the new patient instructions appeared to reduce blocking in causal ratings, there was nonetheless evidence of a reliable blocking effect within the new patient group, $BF_{10} = 12.46$, $d = 0.241$, $t(152) = 2.98$, $p = .002$.[8] With the exception of an interaction between experiment and test order ($BF_{incl}$ = 6.401, $\eta_p^2$ = 0.024, $F(1,308)$ = 7.45, $p = .007$)[9], all interactions with experiment and/or test order were not supported by the data, smallest $BF_{excl}$ = 2.45 ($\eta_p^2$ = 0.004, $F(1,308) = 1.19$, $p = .277$) for the four-way interaction. Finally, there was also a main effect of instruction reflecting higher causal ratings in the new patient group on average, $BF_{10}$ = 2816, *error%* = 0.897, $\eta_p^2$ = 0.066, $F(1,308) = 21.72$, $p < .001$.

[8] Consistent with the approach taken in previous experiments the prior for the alternative hypothesis specified that causal ratings for blocked cues should be lower than for overshadowed cues. Ratings were collapsed over experiment and test order given the lack of any interactions with cue and instruction group.

[9] The experiment x test order interaction indicates that when the two test judgements were separated into distinct test blocks in Experiment 2.3a, average causal ratings were lower when participants made causal ratings first than when they completed a block of memory judgements first. However, this was not the case in Experiment 2.3b, where the order of the test questions on each test trial did not appear to affect average causal ratings. As there were no further interactions with cue type or instruction group, this does not appear to have implications for the observed blocking effect per se, nor the sensitivity of blocking to instructions.

**Figure 2.9**

*Causal ratings for Experiments 2.3a and 2.3b*



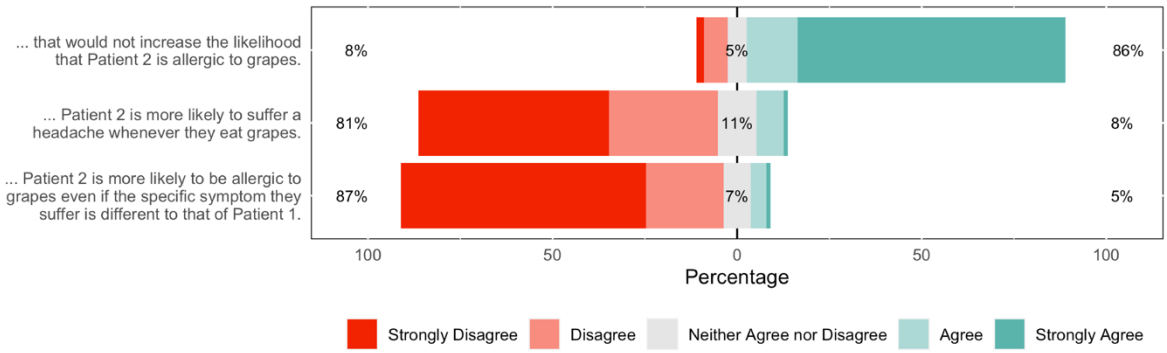

*Note*. Left panels show causal ratings scores from Experiment 2.3a, right panels show the same from 2.3b. The top panels show mean (+/- SEM) causal ratings for the blocked and overshadowed cues. In the bottom row, the distribution of blocking scores (causal ratings for overshadowed - blocked cues) is represented by the shaded bean, with individual blocking scores indicated by dots. The mean and standard error of these distributions are given by the horizontal bar and surrounding shaded box respectively. Ratings and Scores are represented as a function of instruction group (same patient vs. new patient).

**Survey**

Figure 2.10 shows the three survey items on the left and the percentage of participants who agreed or disagreed with those items. Given the scenario that a hypothetical patient (Patient 1) was known to suffer from headaches after eating grapes, a large majority of participants (87%) agreed that "If Patient 1 suffers a headache whenever they eat grapes, that would not increase the likelihood that Patient 2 (another patient) is allergic to grapes." Consistent with this, most participants *disagreed* with the assertion that if one patient suffers headaches when they eat grapes it is more likely that another patient will suffer a headache (81%), or any other symptom of allergic reaction (87%), when they eat grapes.

**Figure 2.10**

*Survey of food allergy assumptions*




*Note.* Participants were asked to indicate on a Likert scale the extent to which they agreed with each of the three statements on the left. The statements all began with "If Patient 1 suffers a headache whenever they eat grapes…"). The figure shows the percentage of participants who agree (right panel) or disagree (left panel) with each statement. Neutral values are shown in grey.

**Discussion**

Experiments 2.3a and 2.3b replicated the primary findings of Experiments 2.1 and 2.2. Introducing a new patient in the training phase of a blocking design appears to attenuate blocking in causal judgements. This observation did not rely on a difference in the base rate of allergies in the pretraining and training phases. The base rates of allergic reactions in the training phases were matched in Experiment 2.3b, such that if causal judgements for the blocked cues reflected the base rate of allergies of the patient under examination, both the same patient and new patient groups should give similar ratings to the blocked cues. However, this was not the case as there was a clear reduction in blocking as a result of the new patient instructions in both experiments. Experiments 2.3a and 2.3b also lend support to the finding that participants' overt predictions made during training were affected by the critical instructions. Accuracy for the blocking compounds was more similar to that for the novel overshadowed compounds in the early stages of training in the new patient group than the same patient group even though both groups had learned the pretrained contingency. This indicates that participants were less likely to make predictions based on what they had learned in pretraining if they were told that the pretraining concerned a different patient.

Although the instructions affected the overt predictions made during training and the causal judgements made at test, Experiments 2.3a and 2.3b were consistent with the earlier experiments in that they had little influence on how well the blocked cue-outcome association was learned. Both the same patient and new patient groups showed poorer retrieval of the outcomes associated with the blocked cues than the overshadowed cues. The differential sensitivity of associative memory and causal judgements does not appear to rely on the order in which these judgements are made, as the evidence did not support a test order effect on either test judgement.

As discussed, the lack of sensitivity to the instructions in learning scores may reflect a lack of cognitive control over the learning system, however an alternative explanation is that the manipulation does not effectively reduce the relevance of pretraining. This may be due to participants assumptions that food allergies tend to follow specific patterns, for example, many

people have allergies to shellfish. It might be reasonable then to assume that if one patient has a specific food allergy, then the next patient will also. However, the results of the survey do not support this explanation of the learning scores as responses reflected the assumption underpinning the critical manipulation in this series of experiments. That is, a large majority of participants endorsed the statement consistent with the assumption that the introduction of a new patient at training should render the information acquired during pretraining largely irrelevant. Specifically, participants agreed that a patient's specific food allergy has little bearing on the likelihood that a new patient will suffer the same allergy.

**Bayesian mini meta-analysis**

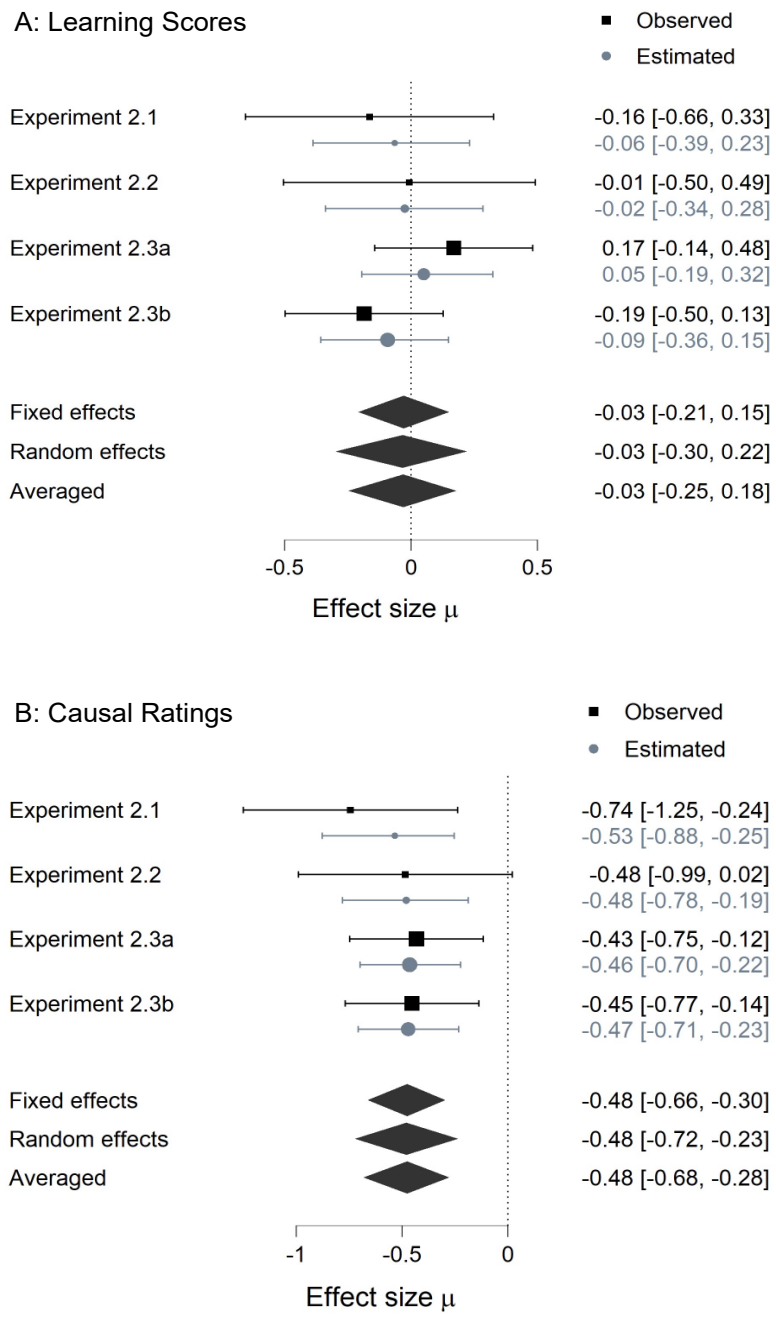
The primary aim of this chapter was to examine the claim that non-associative knowledge could affect learning directly by modifying the predictions that inform prediction-error driven associative learning. To this end, I introduced an instructional manipulation designed to disrupt cue competition by reducing the relevance of the pretraining phase to predictions made during training. The key finding of interest in these experiments was that while the instructions affected assessments of causality, learning scores appeared to be less sensitive to the instructions. However, the analyses did not always provide clear support for this pattern of results, potentially due to a lack of power in the smaller studies (2.1 and 2.2). To quantify the evidence for this pattern of results across the present experiments I performed two meta-analyses (one for each test measure) using the Bayesian model averaging approach (Gronau et al., 2017, 2021). This approach produces an estimate of the effect size pooled across the experiments and provides a measure of the evidence for an effect that considers both random and fixed effects models. In other words, these analyses provide a test of the evidence that the same patient/new patient manipulation affected the magnitude of blocking in learning scores and separately, in causal ratings. Analyses were conducted with the metaBMA package for R (Heck et al., 2019) and the JASP software (JASP Team, 2020).

Effect sizes (Cohen's d) were calculated from independent samples t-tests comparing the size of the blocking effect (overshadowed – blocked) across the two instruction groups for each test

measure within each experiment. The observed effect sizes and standard error for each study are given on the right-hand side of Figure 2.11. A negative effect size indicates that the new patient instructions effectively reduced the magnitude of blocking relative to the same patient instructions. Following the recommendations of Gronau et al. (2021) an Inverse-Gamma prior (1, 0.15) was specified for between-study heterogeneity ($\tau$) and a default Cauchy prior (centred on zero, r = .707) for the overall effect size ($\mu$).

**Figure 2.11**

*Bayesian model-averaged meta-analyses of the effect of instructions on blocking*




*Note.* Forest plot of observed (square marker) and estimated (circle marker) effect sizes (Cohen's d with the 95% credible interval) for the effect of instructions on the magnitude of blocking in learning scores (Panel A) and causal ratings (panel B) for each experiment. The effect size markers are scaled to reflect the relative sample size. Diamond markers represent the mean meta-analytic effect size (μ) from each model and the averaged model.

Figure 2.11 A shows the observed and estimated effect sizes for the blocking effect in learning scores for the four experiments, as well as the meta-analytic estimates of mean effect size for the fixed, random, and model-averaged analyses. The model-averaged Inclusion Bayes Factor for the meta-analytic effect indicated support for the null that the instructions did not affect the magnitude of blocking in learning scores, $BF_{10}$ = 0.12. The mean of the model averaged meta-analytic effect size was -0.031. As Figure 2.11 B shows, the Bayesian meta-analysis of the same effect in causal ratings provided strong support for a moderate effect of instructions on the magnitude of blocking. The Bayes Factor indicated that the data are 83.8 times more likely under the alternative hypothesis that the instructions affected the magnitude of blocking than the null. The model averaged analysis produced a mean meta-analytic effect size estimate of -0.48.

## General Discussion

Three experiments tested whether associative memory encoding is subject to control by instructions in a similar manner to explicit judgements of causality. To achieve this, I introduced instructions designed to either support or undermine competition for association by manipulating the relevance of the associative history of the competing cue in a blocking design. I found that overt predictions made during training were sensitive to instructed control. Similarly, instructions that reduced the relevance of pretraining, also reduced the effect of that pretraining on participants' judgements of causality in a later test phase. That is, causal ratings for the blocked cues were more similar to those for the novel overshadowed cues when participants were instructed to treat the training phase as a new patient, than when the training phase was described as a continuation of pretraining. Although the evidence did not strongly favour the hypothesis in Experiment 2.2, I subsequently replicated the pattern in Experiment 2.3 and the mini meta-analyses are consistent with a moderate effect of instructions on the expression of blocking in causal ratings.

Unlike overt predictions and causal judgements, associative memory was notably resistant to instructed control. I consistently observed blocking in learning scores, the magnitude of which did not interact with instructions about the relevance of pretraining. Thus, although the instructions

successfully reduced the perceived relevance of pretraining (evident in both causal judgements and overt predictions), they had little effect on how the cue-outcome relationships were encoded in memory. In Experiment 2.2, eye gaze data revealed that when participants were making active predictions about the outcome of a trial, there was evidence of learned biases in attention away from blocked cues regardless of instruction group. Similarly, there was a robust blocking effect in selective attention on negation trials that was consistent with the lack of sensitivity observed in associative memory retrieval.

It should be noted that it is difficult to determine what processes contribute to overt eye gaze. For instance, they may accurately reflect learning-related biases in attention (i.e., gaze indicates cue associability) or may instead merely reflect participants' active decisions as part of providing an overt prediction on a given trial. Consistent with the latter, gaze patterns in this experiment largely followed patterns of explicit predictions and of choice on negation trials. However, if gaze does indeed reflect learned attentional biases but is also able to be brought under instructed control in the same way that causal judgements are, how can this be reconciled with the insensitivity of associative memory? One observation that may help us answer this question is the relative stability of the selective attention patterns across the training phase. While there is some evidence for sensitivity to the instructions in the first block of training, by the end of training any sensitivity appears to be lost and overall, there was a robust bias in attention in both groups. Another consideration is the evidence that learning that a cue is predictive in one context, produces relatively automatic changes in attention that carry over into new contexts in which all cues are equally predictive (Le Pelley et al., 2013). These rapid changes in attentional capture would not be well captured by the measures used in this task. Notably these changes also occur in the blocking effect (Luque et al., 2018), and are resistant to instructed control (Cobos et al., 2018). On balance, the evidence is consistent with the notion that selective attention, like associative memory, is relatively difficult to modify with instructions that successfully control judgements of causality.

Past findings have shown that causal judgements are sensitive to information that moderates the relevance of associative memory to assessing causality. Researchers have used instructions, explicit training, or some combination of the two, to successfully modify assumptions about the direction of causality (e.g., Waldmann & Holyoak, 1992) or assumptions about how causes interact to produce an effect (e.g., Lovibond, et al. 2003). The causal judgement data reported here are clearly consistent with these observations but demonstrate this sensitivity in a simple and novel way. It is perhaps surprising then that associative learning was apparently resistant to control by these same instructions. On the one hand, this result is inconsistent with Thorwart and Livesey's (2016) suggestion that the latent predictions derived from past experience with cues might be brought under cognitive control. The instructions, which did have an impact on how associative memory informed causal judgements, had little impact on how those memories were laid down. The persistent blocking in associative memory suggests that the instructions about the relevance of pretraining about A did not appear to modify the magnitude of prediction error experienced when later learning about the novel compound AB. On the other hand, resistance to instructed control may be the result of changes in the associability of cues during pretraining producing biases in attention that are not easily brought under control via instructions. The lack of sensitivity to instructions in the proportion of time spent looking at the target cues is consistent with this idea. These two explanations, resistance to control of prediction error and resistance to control of encoding, are not mutually exclusive. A reasonable conclusion to draw is that both processes must be somewhat resistant to instructed control; if either were strongly affected by the instruction then we should have seen a modulation of the magnitude of the blocking effect in learning scores. Nonetheless it is important to note that even with the addition of eye tracking, it is difficult to tease apart the separate influences of prediction error and selective attention, largely because attention and learning processes themselves are so intimately linked.

Thorwart and Livesey (2016) argued that non-associative knowledge (such as beliefs established via instructions) could impact associative learning via the translation of associative

memory to behaviour. That is, beliefs about the relevance of past associations could result in associative retrieval having a relatively strong or relatively weak influence on variations in responses, and retrieved associations might even be replaced completely in some circumstances by the impact of a reasoned belief. It seems entirely plausible that this level of flexibility to engage (or not engage) with associative memory is available in the case of causal judgements. Of course, just as causal ratings are hypothesised to reflect a translation of associative memory to a behaviour, so too are the learning scores. These test questions were designed to ask much more directly about the strength of the learner's memory for which outcomes were paired with which cues. One can ask, why are ratings of associative memory insensitive to instructions that appear to successfully modify causal judgements? It is not simply the case that causal judgments are controlled by reasoning and inference while other learned behaviours are not. For instance, the resistance to instructed control is inconsistent with evidence from fear conditioning paradigms that giving participants explicit instructions that alter their overt expectancies about receiving a shock can modify conditioned responding to a shock-associated cue in a manner that is at odds with the associative history of that cue (McNally, 1981; Lovibond, 2003). On the other hand, reverse associations such as the Perruchet effect (McAndrew et al., 2012; Tran, et al., 2020; Weidemann et al., 2016) suggest that many learned influences on responding are not entirely controlled by the conscious beliefs of the individual.

Some have argued that when participants are given clear and explicit reasons to use another form of simple inference, it provides a situation in which causal judgments are less likely to reflect associative strength. For instance, this seems to be the case when they are encouraged to use deductive reasoning (Livesey et al., 2019). The current study contributes further evidence to this hypothesis by showing that associative memory and causal judgments diverge when participants are given an explicit reason to disregard associative history. The important observation in this case is that the strength of associative memory (and to a large extent the patterns of overt stimulus encoding) are unaffected by that same instruction. This suggests that learners possess greater flexibility with how they interpret the associations they have learned, than they do over the learning

itself. This naturally raises the question whether and in what circumstances a learner *will* rely on associative retrieval to make causal judgments. As the discussion above suggests, one possibility is that this is the default mode that people engage in when they have no strong impetus to reason in other ways. It is also a possible that a tendency to make judgments based on associative memory is something that varies among individuals (Goldwater et al., 2018). Although these are important questions, they are beyond the scope of this thesis. Instead, Chapters 3 and 4 provide further tests of the hypothesis that instructions can modify the way associative memories are encoded.

# CHAPTER 3

## Instructed known causes in the blocking effect

I have argued that the strength of associative memory, or the relative ease with which an outcome is brought to mind in the presence of a cue, is sometimes interpreted as evidence for or against a causal relationship between the two events. This may be especially true when there is a lack of other evidence on which to judge causality or when careful reasoning about causal structure is not actively encouraged. As demonstrated in the previous chapter however, beliefs about the relevance of associative history can influence the extent to which causal judgements reflect the strength of associative memory. That is, causal judgements and the overt predictions made during training clearly diverged from learning scores in their sensitivity to instructions. The expression of blocking in causal judgements was reduced when the task instructions reduced the relevance of the pretraining of the blocking cue. Similarly, the overt predictions made early in training reflected the influence of the instructed relevance of pretraining. In contrast, blocking persisted in learning scores regardless of how relevant the pretraining information was implied to be.

Many accounts of learning assume that differential prediction error should be critical for establishing the blocking effect; learning about the blocked cue is reduced because the outcome is already well predicted (see Livesey, 2023, for a recent summary). Assuming that prediction error *is* a critical factor in determining the strength of learning, the results of the previous chapter imply that successfully modifying causal judgements and even the overt predictions made by participants during training does not necessarily have a strong impact on prediction error learning. In other words, the predictions important for forming associative memories may not be easily brought under control by instructions that successfully modify overt predictions and judgements.

One explanation for the lack of sensitivity to instructions is that the instruction manipulation employed in Chapter 2 modified predictions indirectly, and perhaps not wholly effectively, by

introducing a patient change between phases. This should reduce the relevance of the pretraining to the generation of predictions during the training phase and indeed the test of causal judgements reflects this assumption. However, as discussed in the previous chapter, participants may assume that the same foods control allergic reactions in both contexts as certain foods, like peanuts for example, are more commonly associated with allergies outside the laboratory. This criticism has been leveled at experiments that used similar instructional manipulations in the allergist task to investigate the learned predictiveness effect (Mitchell et al. 2012). For instance, in a classic demonstration of learned predictiveness, foods that were previously predictive in one context were learned about more readily than previously nonpredictive foods even though all foods in the new context (instantiated by instructions indicating that a change in patient had occurred) were equally predictive (Le Pelley & McLaren, 2003). Mitchell and colleagues (2012) argued that this is not sufficient evidence of an associatively acquired bias as participants could simply deduce that one patient's allergies are informative for the next patient's allergies, given the regularities in food allergies in general. Experiment 2.3b addressed this criticism by surveying participant's assumptions about the relevance of one patient's allergies to another's; when asked explicitly, most participants do not endorse the belief that the observation of one patient's allergies changes the probability that another patient will suffer an allergy to the same food.

It is possible, however, that associative encoding is affected by instructed changes in predictions but there are further constraints on when such an influence might be apparent. For instance, it could be that for instructed changes to a learner's beliefs to have an influence on associative encoding, greater effort (e.g., attentional control) needs to be applied than to influence causal judgments. Since the switch from same patient to new patient brings with it an increase in uncertainty, the learner may be unwilling to apply this effort. With this in mind, this Chapter addresses a complementary question, whether

instructed changes in prediction will influence learning scores when a more direct instruction brings greater certainty about the causal status of (some of) the cues.

The learned predictiveness literature provides two examples of how this can be achieved. First, Mitchell and colleagues (2012) adopted a reversal instruction, that is they directly instructed participants that those foods that had previously been predictive in one context would no longer be so, and previously non-predictive foods would now be predictive. Contrasting this with a group that received instructions that the relationships that held in the previous phase would continue into the new context, they found that reversal instructions were sufficient to reverse the learned predictiveness effect in causal judgements. However, Shone and colleagues (2015) were unable to replicate this result and instead found no bias in learning following reversal instructions. This is perhaps unsurprising as a reversal instruction is inherently more difficult to apply than its counterpart, the continuity instruction (for example, under reversal instructions, one arguably has to first select the cues established as predictive in the first phase, then re-orient attention to those cues that were previously non-predictive). Thus, if the procedure fails to produce a full reversal in responding this could be attributed to the nature of the instruction, rather than a difference in learning or causal reasoning. The second approach was applied by Shone et al. (2015) in response to the failure to replicate the reversal of the learned predictiveness effect in their hands. In this instance, participants were provided with explicit instructions that directly identified the foods that would be predictive in the new context. That is, following training in which one set of foods was established as reliably predictive of allergy in one patient, participants were explicitly told which foods the new patient was allergic to. Half of the foods identified as allergens by the instructions were previously predictive foods (congruent) and the remaining half were previously non predictive foods (incongruent). This procedure removes the confound in the previous approach, thus any differences in learning observed must be the result of the biases established in Phase 1, rather than a difference in the difficulty of applying the instruction. They found that explicit instruction about

the known causes of allergies was sufficient to reverse the learned predictiveness effect in causal judgements. Yet, the effect of learned predictiveness among foods identified by instructions as known allergens persisted such that participants learned more about previously predictive foods than previously nonpredictive foods.

The following experiments employ the second of these two approaches, explicit instructions that identify the known allergens of the new patient, to investigate the possibility that a more direct set of instructed beliefs — ones that do not increase the uncertainty of the learner but rather replace one set of confident beliefs (about pretrained causes) with another (about instructed causes) — lead to a stronger influence of instructions on associative encoding.

As in the experiments in the previous chapter, participants in these experiments were asked to complete an allergist task. Consistent with the new patient groups in Chapter 2, all participants received instructions that indicated that the pretraining and training phases referred to two separate patients. As such, all else being equal, I expected to see clear evidence of blocking in learning scores, which is reduced or absent in causal ratings. In both experiments in this chapter, the cues of greatest interest were blocked and overshadowed food cues that were each paired with a cue from a specific food category (a dairy or meat product). For blocked cues, their competing cue was pretrained with an allergic reaction outcome whereas for the overshadowing cue, the competing cue was novel at the beginning of the learning phase. In both experiments, two groups of participants received identical sets of contingencies, using the new patient condition from the previous experiment as the default cover story. The key instruction manipulation involved informing one group of participants about specific allergic reactions suffered by the new patient introduced in the learning phase. By introducing these known causes, it was assumed that predictions based on the prior associative history of the pretrained cues would have less impact on learning about the blocked cues, rendering them more similar to the equivalent overshadowing cues. To test this, the primary analyses in these experiments concerned the

learning scores for blocked and overshadowed cues, with the hypothesis that the addition of instructed known causes competing with these cues would reduce an otherwise robust blocking effect. However equivalent analyses of causal ratings are also reported and discussed where relevant. The designs also provide opportunities to test for other effects of instruction-based predictions on learning and cue competition, which will be discussed in relation to the individual experiments.

## Experiment 3.1

Experiment 3.1 compared two groups who received different instructions at the beginning of the learning stage, based off the new patient instruction conditions of the previous chapter. The *neutral* group was identical to the new patient groups from the previous chapter in that they were given no other information about the new patient's allergies. The *known cause* group on the other hand were told that the new patient was already known to suffer from a set of specific food allergies. This experiment builds on the relevance manipulation employed in Chapter 2 by giving one group more explicit instructions on what to expect in the training phase, which provide certainty about the causal status of a subset of the cues. By instructing the known cause group that the new patient is known to suffer from a specific set of food allergies, I am giving this group an alternative set of information which should be considered more relevant than anything learned in pretraining. The prediction was thus that this group would be less guided by previously learned associative predictions than the neutral group, who may rely on prior predictions despite believing that they are not particularly relevant to the current patient.

The bolded letters in Table 3.1 indicate which foods were identified in the known cause group as known allergies of the training phase patient. In a standard blocking design, the appropriate control cue for a blocking effect would be one that was only ever trained in a compound with another novel cue (an 'overshadowed' control cue). With such a control we can be relatively sure that any decrement in

learning score and/or causal judgement for the blocked cue is due to the prior training of the pretrained cue. That is still the case in the neutral group, but this would be a poor control cue for the known cause group as the pretrained cues (A and C) are subject to both prior training *and* direct instruction about which allergic reaction is likely to occur. To address this, one of the cues from each overshadowing compound (E and G) was described as a known allergy and it is their competitor cues not subjected to instruction (F and H) that were used as appropriate controls for the blocked cues (B and D). Of course, this was only a concern in the group that received explicit instructions about the known allergies of the new patient and therefore the distinction between E and G and F and H is meaningless for those in the neutral group who did not receive these instructions. However, for consistency, F and H served as the control cues for both instruction groups.[10]

Another notable difference between the design of this experiment and those employed in Chapter 2, is the absence of the no allergic reaction outcome. The design of Experiment 3.1 was developed independently of those in the previous chapter and so the aim of Experiment 3.2 was to replicate the results of this experiment with a design that employed a 'no allergic reaction' outcome. Two aspects of the design in the present experiment carry implications for the logic of the task. The first of these is the modification of the unovershadowing contingencies. Rather than pretraining a non-allergenic food, the unovershadowing contingencies involved a change in the symptom of allergic reaction from pretraining to training. Experiencing a change in symptom when a new food was added to the meal could confuse participants if the examples were derived from the same patient, as experienced by same patient groups in Chapter 2. However, in this instance all participants received a patient change between phases. Second, removing all instances of non-allergenic outcomes affects the base rate of allergic reactions. In this design the patient always suffers an allergic reaction but the specific symptom

[10] Bayesian paired samples t-tests comparing responses for E/G to F/H for those in the neutral group confirmed that there was no difference in learning scores, $BF_{01}$ = 3.308, *error%* = 0.005, in favour of the null, for causal ratings the evidence for a difference was equivocal, $BF_{10}$ = 1.549, *error%* <.001.

they experience differs depending on which food item or meal was consumed. As discussed in Chapter 1, modifying the base rate of allergic reactions in a blocking design can affect the strength of the blocking effect observed in causal ratings. In this instance, although the causal status of the blocked cue remains ambiguous, increasing the base rate of allergic reactions should reduce uncertainty about the likelihood that blocked cue is causing an allergic reaction. However, this should also be true of the overshadowed control cues.

**Table 3.1**

*Design of Experiment 3.1*

| | Pretraining | Training | Test |
|---|---|---|---|
| Blocking | A – O1 | **A**B – O1 | A, B |
| | C – O2 | **C**D – O2 | C, D |
| Overshadowing | | **E**F – O1 | E/F |
| | | **G**H – O2 | G/H |
| Unovershadowing | I – O1 | **I**J – O2 | I, J |
| | K – O2 | **K**L – O1 | K, L |
| Fillers | | M – O1 | M |
| | | N – O2 | N |
| | OP – O1 | **S** – O1 | S |
| | QR – O2 | **T** – O2 | T |

*Note.* O1/O2 denote the different allergic symptoms that the patient experienced (randomly assigned to fever /nausea). Letters (A – T) represent the different food cues, the known allergies of the training phase patient are indicated by bold type. For example, participants in the *known cause* group were instructed that the training phase patient suffered from Outcome 1 (O1) after eating food **A**.

## Method

### Participants

Sixty-four University of Sydney undergraduate psychology students participated in exchange for partial credit. One from the neutral group was excluded for failing to reach 60% accuracy in the final block of training phase, leaving a final sample of 63 participants (49 female, *M* age = 18.73). Participants were randomly assigned to the neutral (*n* = 31) and known cause instruction (*n* = 32) groups.

### Apparatus and Stimuli

Twenty food items served as cues A –T: *Yoghurt, Cheese, Milk, Ice cream, Bacon, Beef, Chicken, Fish, Lemon, Garlic, Bread, Peanuts, Avocado, Corn, Mushrooms, Apple, Banana, Pasta, Rice,* and *Carrots.* Of these, eight foods served as the instructed allergies for the patient from the training phase (See bolded cues in Table 3.1) four dairy foods and four types of meat, the remaining 12 were foods from any other category.

### Procedure

The procedure was similar to that employed in Chapter 2. Participants were asked to play the role of a doctor attempting to discover the allergies of their patients. On each trial, participants were presented with one or two foods that the patient had consumed and were asked to predict which allergic reaction would follow. They then received corrective feedback before continuing to the next trial. Pretraining proceeded for five blocks, with each block containing two iterations of the trial types listed in Table 3.1.

At the end of pretraining the critical instructions were introduced. All participants were told that they would be seeing a new patient, Miss Y, in the next phase (training), those in the neutral group received the same instructions as the new patient groups in Chapter 2. That is, they were given no specific information about the new patient's allergies. Those in the known cause group, on the other hand, were given explicit instructions regarding the allergies of the new patient. The instructions

outlined eight known allergies of the new patient (bolded cues in Table 3.1). To increase retention of this information the instructed food allergies were drawn from two categories of food items, dairy or meat, and each category was associated with just one type of allergic reaction. The instructions gave participants both the category of foods that caused a specific reaction in the new patient, as well as the names of individual foods within that category that they would see in the next phase. For example, those in the known cause instruction group received the following instruction:

> The following information is ALREADY KNOWN about Miss Y's allergies: Miss Y suffers from NAUSEA after eating DAIRY such as YOGHURT, CHEESE, MILK, and ICECREAM
>
> AND
>
> Miss Y suffers from FEVER after eating MEAT such as BACON, BEEF, CHICKEN, and FISH

Training trials proceeded in a similar manner to pretraining, although training was limited to two blocks (rather than the three blocks of training in earlier experiments) to minimize ceiling effects in the known cause group. Participants therefore saw each trial type from training four times in total.

Consistent with the procedure of experiments presented in Chapter 2, each cue from training was presented individually in a randomised order during the test phase and participants were asked to recall the outcome that followed the cue and then to assess the causal status of the cue. The memory and causal ratings were intended to separate the inferences participants made about the causal status of the cues from their associative memory for the cue-outcome pairings and this distinction was emphasized in the instructions. Unlike previous experiments, there were only two possible outcomes

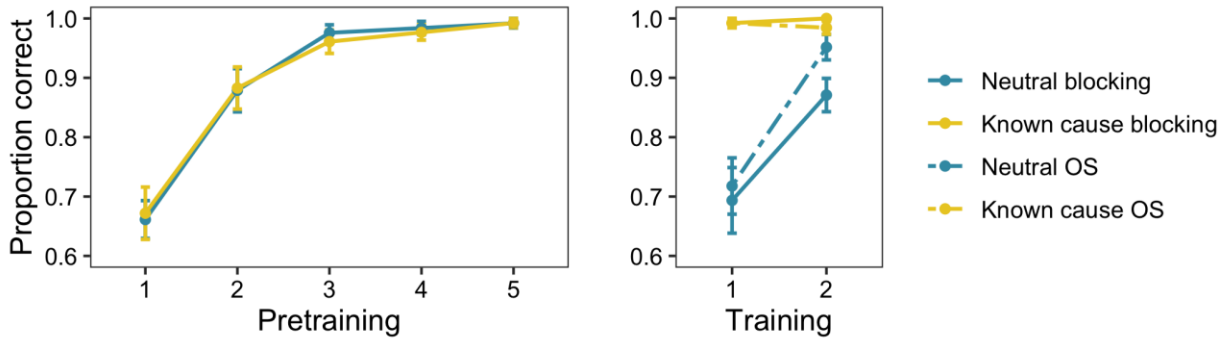
(Outcome 1 or Outcome 2, both allergic reactions) so participants made the recall judgment on a single scale ranging from “definitely paired with O1” to “definitely paired with O2”. Both scales were presented on a single screen, and participants were required to make both judgements before confirming their selections by pressing the spacebar to proceed to the next trial (see Appendix B for an example).

## Results

Figure 3.1 shows the proportion of correct predictions for the blocked and overshadowed contingencies on each trial as a function of instruction group. Both groups demonstrated evidence of learning; accuracy for all critical contingencies was above 98% by the last block of pretraining. As expected, the explicit instructions regarding the allergies of the new patient had a marked effect on predictions during training. In the known cause group, mean accuracy for the compounds containing an instructed cue was close to 99% on average, suggesting that the instructions effectively modified participant’s overt predictions. Performance was substantially poorer in the training phase for the neutral group. As in the previous chapter we examined performance in the first block of training immediately following the critical instruction. A Bayesian compound (blocked vs overshadowed) by instruction (neutral vs. known cause) ANOVA revealed that the instruction had a main effect on accuracy in the first block of training (trials 1 and 2), $BF_{10}$ = 1.313 x$10^6$, *error%* = 2.984, $\eta_p^2$ = 0.423, $F(1,61) = 44.68$, $p < .001$. Overall accuracy was substantially higher following known cause instructions regardless of compound ($BF_{Excl}$ = 3.521, $\eta_p^2$ = 0.003, $F(1,61) < 1$, $p = .669$, against the interaction, and $BF_{01}$ = 4.785, *error%* = 2.621, $\eta_p^2$ = 0.003, $F(1,61) < 1$, $p = .669$ in favour of the null over the main effect of compound).

**Figure 3.1**

*Accuracy of training predictions in Experiment 3.1*




*Note.* Mean (+/- SEM) proportion correct on each block of pretraining and training of Experiment 3.1. Shown here as a function of instruction group (neutral vs. known cause) and cue or compound: blocking (A/C in pretraining, AB/CD in training; labelled A-D) and overshadowing (EF/GH; labelled E-H).

**Test**

Figure 3.2 shows the learning scores and causal ratings for the critical cues in Experiment 3.1. To assess blocking in causal ratings we conducted a Bayesian ANOVA with factors of cue (B/D vs. F/H) and instruction (neutral vs. known cause). There was strong evidence for an overall blocking effect (lower causal ratings for B/D than for F/H), $BF_{10}$ = 275, *error%* = 2.872, $\eta_p^2$ = 0.239, $F(1,61) = 19.14$, $p < .001$. Ratings were also on average higher in the neutral group than the known cause group $BF_{10}$ = 612852, *error%* = 1.144, $\eta_p^2$ = 0.408, $F(1,61) = 42.09$, $p < .001$, suggesting that those in the known cause group were on average more inclined to indicate that the uninstructed cues were not causing the new patient's allergies than those in the neutral group, who were not given any explicit information about the pattern of allergies of the new patient. Evidence for a cue by instruction interaction was equivocal, $BF_{Incl}$ = 1.120, $\eta_p^2$ = 0.055, $F(1,61) = 3.55$, $p = .064$. However, an outlier with a blocking score (higher ratings for F/H than B/D) more than three standard deviations above the group mean was identified in the known cause group. Removing this participant from the analysis yielded the same pattern for the main effects ($BF_{10}$ = 239, *error%* = 1.126, $\eta_p^2$ = 0.247, $F(1,60) = 19.72$, $p < .001$, and $BF_{10}$ = 628870, *error%*

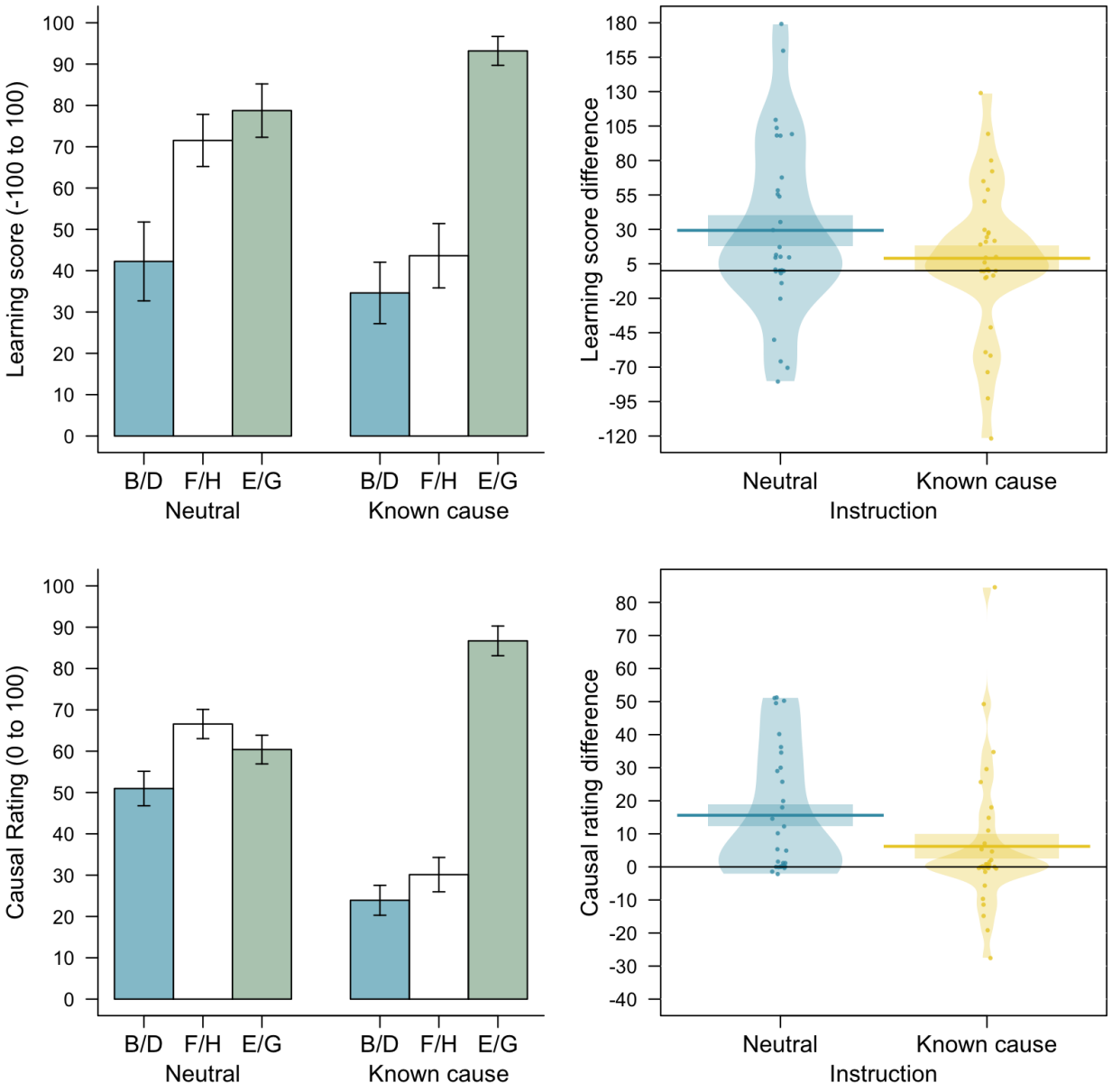
= 2.376, $\eta_p^2$ = 0.415, $F(1,60) = 42.52$, $p < .001$ for the cue an instruction effect respectively) but reduced the ambiguity around the interaction between them. The data were five times more likely under a model that included the interaction, $BF_{Incl} = 5.0$, $\eta_p^2 = 0.112$, $F(1,60) = 7.53$, $p = .008$, reflecting evidence of stronger blocking in the neutral than the known cause group. The results reported below are taken from the sample excluding the outlier; where relevant, the results from the full sample can be found in the footnotes.

To further examine the interaction two Bayesian t-tests compared causal ratings for F/H to B/D in the known cause and neutral groups separately. Consistent with the analysis in previous chapters, a directional alternative (F/H > B/D) was specified for both tests. In the known cause group, the evidence for or against a blocking effect was weak, $BF_{01} = 1.308$ in favour of the null, *error%* < 0.001, $d = 0.239$, $t(30) = 1.33$, $p = .097$.[11] The test of blocking in the neutral group also served as a test of the replication of the new patient conditions from Chapter 2. If the neutral group replicated previous experiments, we would expect to see evidence for the null. Instead, consistent with the evidence from the ANOVA, there appeared to be strong support for a blocking effect in causal ratings in the neutral group, $BF_{10} = 809$, *error%* < 0.001, $d = 0.837$, $t(30) = 4.66$, $p < .001$. This robust blocking effect appears to be larger than those observed in the same patient groups in Chapter 2, something I will return to in the Discussion.

[11] With the outlier included in the dataset the evidence of blocking in the known patient group remained equivocal, $BF_{01} = 1.267$, *error%* ~ 0.001, $d = 0.298$, $t(31) = 1.69$, $p = .051$.

**Figure 3.2**

*Experiment 3.1 learning scores and causal ratings*




*Note.* Learning scores (top panels) and causal ratings (bottom panels) for the critical cues from Experiment 3.1 as a function of instruction group (neutral vs. known cause). The left two panels plot the means and standard errors for the blocked cues (B/D), and the overshadowed cues split into non-instructed (F/H) and instructed (E/G) elements of the compounds – note that this distinction is only relevant for the known cause group. The right-hand panels show the distributions of the blocking scores (F/H – B/D). Each dot represents an individual participant's blocking score, the mean and standard error are given by the horizontal bar and shaded boxes respectively.

Consistent with the findings outlined in Chapter 2, the analysis of learning scores yielded evidence of an overall blocking effect, $BF_{10}$ = 4.582, *error%* = 0.972, $\eta_p^2$ = 0.101, $F(1,61) = 6.82$, $p = .011$, but the evidence for a main effect of instruction, or for an interaction was equivocal, $BF_{10}$ = 1.213, $\eta_p^2$ = 0.069, $F(1,61) = 4.55$, $p = .037$, and $BF_{Incl}$ = 0.629, $\eta_p^2$ = 0.030, $F(1,61) = 1.91$, $p = .172$ respectively.[12] Planned Bayesian t-tests with informed priors for the alternative in the direction of blocking (F/H > B/D) showed that, like the comparable new patient groups from Chapter 2, the neutral group demonstrated a reliable blocking effect in learning scores, $BF_{10}$ = 6.454, *error%* < 0.001, $d = 0.466$, $t(30) = 2.60$, $p = .007$. In contrast, there was weak evidence to indicate that the known cause group did not show blocking in learning scores, $BF_{01}$ = 2.753, *error%* < 0.001, $d = 0.128$, $t(30) = 0.71$, $p = .482$.[13]

Although it was not the primary aim of this experiment, there was some indication of what I will refer to as an *instructed blocking* effect for those participants given explicit instructions about the known allergies of the new patient. The overshadowed cues E – H, which were novel at the beginning of training, provide a test of the effect of targeted instructions on cue competition. That is, informing participants that one cue in a novel compound causes the outcome they observe may produce an instructed blocking effect that is comparable to the kinds of cue competition effects observed following explicit pretraining of a predictive cue. Two complementary analyses were conducted to explore this. The first compared the instructed to uninstructed elements within the overshadowing compounds (EF & GH). This gives some indication of the effect of the known cause instructions on the acquisition of instructed and competing cues within a compound but is not an appropriate test of blocking, *per se* (a limitation addressed in the next experiment). Two Bayesian t-tests compared learning scores, and separately causal ratings, for the instructed (E/G) to the uninstructed cues (F/H) of the overshadowed

[12] Including the outlier did not alter the pattern of results: $BF_{10}$ = 4.398 *error%* = 0.981 for the cue main effect, $BF_{10}$ = 0.898 *error%* = 1.031 for the instruction main effect, and $BF_{Incl}$ = 0.560 for the cue by instruction interaction.
[13] Including the outlier weakened support for the null but not by a large margin, $BF_{01}$ = 2.128, *error%* < 0.001, $d = 0.169$, $t(31) = 0.96$, $p = .347$.

compounds for those participants in the known cause group. Both t-tests used an informed prior for the alternative that assumes that responses on both measures will be higher for the instructed than the non-instructed elements of the compound (E/G > F/H). The evidence strongly supported the alternative in both learning scores, $BF_{10}$ = 16884, *error%* < 0.001, *d* = 1.048, *t*(30) = 5.84, *p* < .001, and causal ratings, $BF_{10}$ = 2.064 x $10^9$, *error%* < 0.001, *d* = 1.928, *t*(30) = 10.73, *p* < .001[14].

The second analysis assessed whether the additional information about E/G from the known cause instructions affected acquisition of the competing cues F/H by comparing responses to F/H across instruction groups. A Bayesian independent samples t-test with an informed prior assuming scores for uninstructed OS cues would be higher in the neutral instruction group than the known cause group yielded evidence in support of the alternative, $BF_{10}$ = 14.607, *error%* < 0.001, *d* = 0.728, *t*(60) = 2.87, *p* = .003, indicative of stronger competition from cues that are directly instructed as known causes of allergy than from novel (uninstructed) cues. This pattern was also observed in causal ratings, $BF_{10}$ = 1.848 x $10^7$, *error%* < 0.001, *d* = 1.848, *t*(60) = 7.28, *p* < .001[15].

**Discussion**

As predicted, accuracy was very high in the training phase for those informed about the new patient's food allergies. Participants readily applied the known cause instructions to their predictions in training, regardless of whether the meals presented comprised novel foods or contained pretrained allergens of another patient. Similarly, when given targeted and explicit instructions as to which foods cause an allergic reaction in the new patient, participants gave strong causal ratings to cues instructed as known causes and relatively weak ratings to their competing (non-instructed) cues. In both Chapter 2 and the present experiment, participants given 'new patient' instructions were conservative in

[14] Including the outlier yielded the same pattern: $BF_{10}$ = 17326, *error%* < 0.001, *d* = 1.026, *t*(31) = 5.81, *p* < .001 learning scores and $BF_{10}$ = 1.403 x $10^9$, *error%* < 0.001, *d* = 1.835, *t*(31) = 10.38, *p* < .001 in causal ratings.
[15] Including the outlier did not change the pattern of results: $BF_{10}$ = 11.973, *error%* < 0.001, *d* = 0.700, *t*(61) = 2.78, *p* = .007 in learning scores and $BF_{10}$ = 2.081 x $10^6$, *error%* < 0.001, *d* = 1.679, *t*(61) = 6.66, *p* < .001 in causal ratings.

generalizing predictions drawn from the pretraining patient. The strong performance in the known cause group indicates that when an alternative, and presumably reliable, source of information about the new patient's allergies is available, participants appear to update their predictions and judgments of causality in line with the information given in the instructions.

Regarding the blocking effect, informing participants that the pretrained cue was a known allergen of the new patient weakened the blocking effect in causal judgements relative to when instructions provided no specific information about the new patient's allergies. Evidence for blocking in the causal ratings of the known cause group was equivocal, supporting neither the presence nor absence of a blocking effect. This stands in stark contrast to the robust blocking effect evidenced in the neutral group. The lack of evidence for a blocking effect in the known cause group, along with the evidence that the magnitude of blocking was modified by instructions suggests that pretraining did not modify the competition produced by instruction alone in an additive fashion. However, it is important to note that the evidence for this pattern only emerged when an outlier was removed from the known cause group.

The evidence for blocking in the learning scores was again not clear cut. Although there was an overall blocking effect, it is clear from Figure 3.2 that blocking was numerically weaker in the known cause group, but the ANOVA did not yield support for either the presence or absence of an interaction with instructions. On the other hand, the planned comparisons on the blocking effect within each instruction group were more consistent with the pattern indicated by the means. That is, the moderate evidence for the presence of blocking in the learning scores of the neutral group ($BF_{10}$ = 6.454) can be contrasted with the weak evidence in support of the absence of blocking in the known cause group ($BF_{01}$ = 2.753). Given the inconsistencies it is difficult to interpret the effect of the directed instructions on the cue competition observed in learning scores at this stage. I will return to this in the General Discussion.

It is noteworthy that the neutral group produced a robust blocking effect in causal ratings. The experiments in Chapter 2 showed that although introducing a change in patient did not alter competition for associative memory, it did reduce the magnitude of blocking observed in causal judgements. Although there was still support for the presence of a blocking effect in causal ratings in the new patient groups in Experiment 2.3, the blocking effect observed here in the neutral group appears to be numerically larger compared to that of the new patient groups from Chapter 2, taken as a whole. I did not include a same patient control in this experiment and thus cannot judge whether the new patient instructions reduced blocking in causal ratings. Nevertheless, even if they did, it appears they were not as effective in doing so in this experiment compared to those in Chapter 2. This raises the question, why the discrepancy? There are a number of differences in the way this experiment was run compared to those in Chapter 2, including the lack of no allergic reaction trials. It is possible that this omission encouraged participants to treat the task differently, perhaps more like a categorization or memory task (one in which there is always a correct reaction, it is simply a matter of remembering which one). This explanation is speculative and, given the ambiguity of the result, I will refrain from discussing it further here.

Finally, there was some indication of instructed blocking of both learning scores and causal judgements. That is, directly telling participants that one food from a novel compound of two foods was a known allergen appeared to produce weaker associative memory and lower causal judgements for the non-instructed competing foods in the compound than overshadowing alone (as occurred in the new patient group). The most direct evidence for this effect in Experiment 3.1 is a comparison of the uninstructed overshadowing cues (F and H) across the two instruction groups. However, this comparison is potentially confounded by other more general differences in the way individuals respond to cues during test as a consequence of the instructions they received. Other potential differences between the groups in producing the effects cannot be excluded, for instance the known patient instructions reduced

learning scores and causal ratings for non-instructed cues on average, relative to the new patient instructions. Experiment 3.2 aimed to address this shortcoming by including an appropriate control in the design and re-introducing the no allergic reaction trials into the learning task.

**Experiment 3.2**

Given the ambiguity of the results in the previous experiment, Experiment 3.2 aimed to replicate Experiment 3.1 with an improved design. Two modifications were made to resolve some issues with Experiment 3.1 and extend the findings regarding instructed blocking. The first was to include "no allergic reaction" as an outcome in the design, bringing it in line with those employed in the previous chapter. Although there is no in-principle reason to assume that the inclusion of no outcome trials is critical to these effects, I speculated that it is possible that including these trials contributes to the learner treating the task as one involving causal reasoning rather than one primarily related to categorization or memorizing correct answers. The fact that some trials do not lead to an allergic reaction necessarily implies that not all foods are allergenic, and this may encourage the learner to consider the evidence for *and* against a hypothesis that a particular cue is causal.

The second modification was made to provide an appropriate test of what I termed the instructed blocking effect in Experiment 3.1. As discussed, the known cause instructions appeared to be highly effective at modifying participants' predictions during training, accuracy was close to ceiling in the first block suggesting participants readily applied the information about the new patient's allergies to make their predictions. This raises the question of whether this kind of instruction creates cue competition that is comparable to the kind produced by previous learning. That is, do the same deficits in associative memory occur when the predictions that drive competition are derived from instructions? Two new compounds were added to the design to serve as control cues for the instructed blocking effect, labelled 'overshadowing non-instructed' in Table 3.2. These compounds were presented in the training phase and comprised two novel cues per compound, paired with each of the allergic reactions,

drawn from a pool of foods that were not instructed to be allergenic to the training patient. In other words, they correspond to the standard overshadowing compounds employed in a blocking design, subject only to competition from each other and not from previous learning or explicit instruction. If the known cause instructions produce an effect akin to blocking, I would expect to see poorer associative memory for the cues paired with an instructed known cause of allergies in the new patient (F/H) than for the overshadowed cues that contained no instructed cue (M-P).

## Method

### Participants

Sixty-four undergraduate psychology students from the University of Sydney participated in exchange for partial course credit (48 female, *M* age = 19.08). Participants were randomly assigned to the neutral (n = 32) or known cause (n = 32) instruction groups. All participants passed the training criterion applied to Experiment 3.1.

### Procedure

This experiment employed the same procedure as Experiment 3.1 with the following modifications. A third outcome “no allergic reaction”, as well as four new food items (*Olive oil, Strawberries, Peas, and Broccoli),* were added to the stimulus set. Due to the addition of the “no allergic reaction” outcome to the design, the memory judgements could not be made on a single scale as they were in Experiment 3.1. For this reason, the test procedure was changed to match that used in Chapter 2. Participants were asked to make three separate memory ratings for each cue, indicating their confidence that the cue was followed by each of the three possible outcomes on a scale ranging from “Always followed this food” to “Never followed this food.” The contingency design can be found in Table 3.2.

**Table 3.2.**

*Experiment 3.2 Design*

| | Pretraining | Training | Test |
|---|---|---|---|
| Blocking | A – O1 | **A**B – O1 | A, B |
| | C – O2 | **C**D – O2 | C, D |
| Overshadowing instructed | | **E**F – O1 | E, F |
| | | **G**H – O2 | G, H |
| Unovershadowing | I – no O | **I**J – O2 | I, J |
| | K – no O | **K**L – O1 | K, L |
| Overshadowing non-instructed | | MN – O1 | M, N |
| | | OP – O2 | O, P |
| Fillers | | **Q** – O1 | Q |
| | | **R** – O2 | R |
| | | ST – no O | S, T |
| | UV – O1 | | |
| | XY – O2 | | |

*Note:* O1/O2 denote the different symptoms that the patient experienced (randomly assigned to fever /nausea), 'no O' denotes "no reaction". Letters (A – Y) represent the different food cues, those in bold are the cues that were described as known allergens for participants in the *known cause* instruction group. For example, participants in the *known cause* group were instructed that the patient from the training phase suffered from Outcome 1 (O1) after eating food **A**.

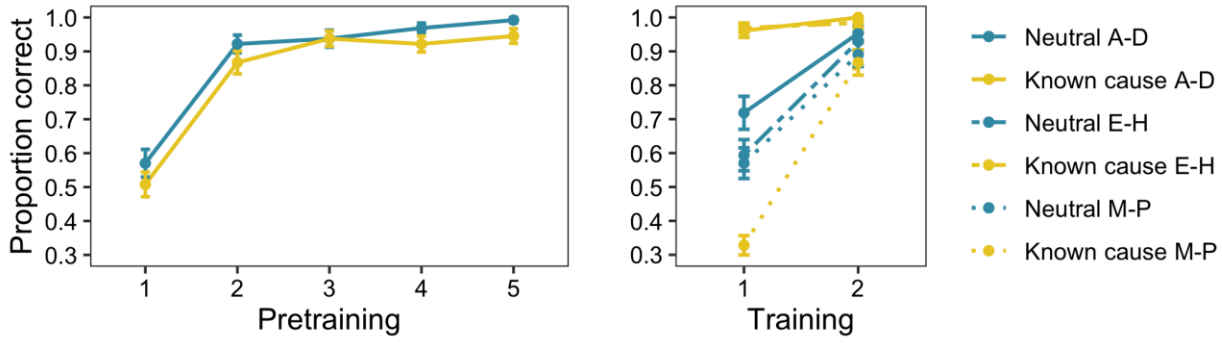
## Results

Figure 3.4 shows the accuracy of predictions made during training for the critical contingencies. By the end of pretraining, accuracy was above 90% for all critical contingencies. Again, there was a clear divergence in performance during the training phase for the instruction groups suggesting that the

instruction effectively modified overt predictions about the allergies of the new patient. This is supported by the analysis of performance in the block immediately following the critical instructions where there was a main effect of instructions, with higher accuracy on average when the patient's allergies were known compared to the neutral condition, $BF_{10}$ = 8.85 x $10^7$, *error%* = 1.085, $\eta_p^2$ = 0.508, $F(1,62) = 64.00$, $p < .001$, The evidence for an effect of compound (blocking (A-D) vs. overshadowing (E-H)), $BF_{10}$ = 0.730, *error%* = 0.893, $\eta_p^2$ = 0.050, $F(1,62) = 3.26$, $p = .076$, and of an interaction between compound and instruction, $BF_{Incl}$ = 1.593, $\eta_p^2$ = 0.063, $F(1,62) = 4.18$, $p = .045$, was equivocal.

**Figure 3.4**

*Training performance from Experiment 3.2*




*Note.* Mean proportion correct responses during pretraining (Blocks 1-5) and training (Blocks 1-2). Line colour indicates instruction group (neutral/known cause). Solid lines show performance for blocking cues and compounds (A/C in pretraining, AB/CD in training), dashed lines for overshadowing compounds with an instructed element (EF/GH), and dotted lines for overshadowing control compounds (MN/OP). Error bars represent the standard error of the mean.

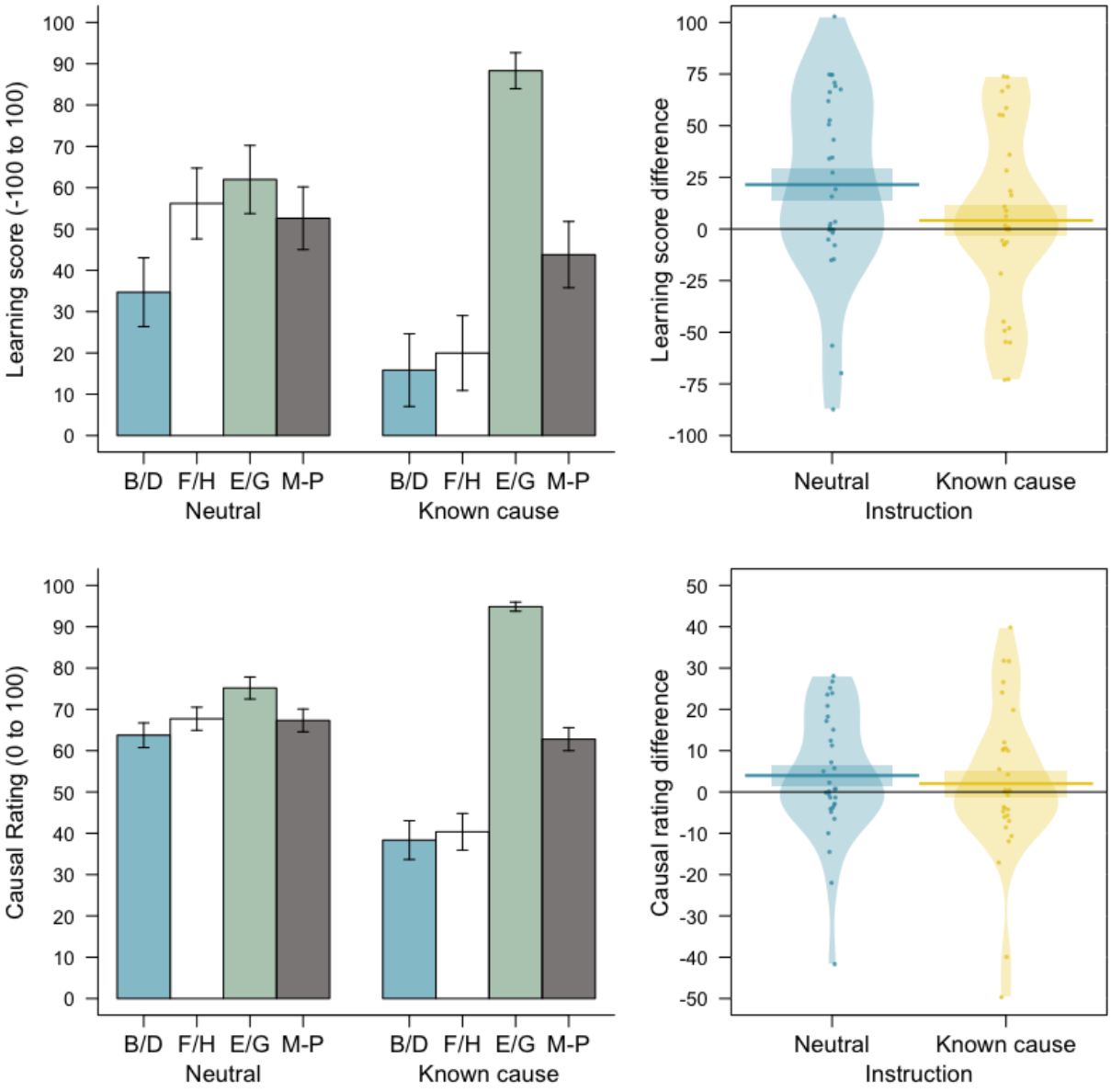
**Test**

Learning scores and causal ratings for the critical cues are given in Figure 3.5. Following the previous experiment, blocking was assessed by comparing responses for the blocked cues (B/D) to those for the overshadowed cues that were paired with a directly instructed cue (in the known cause group) or another novel cue (in the neutral group) during training (F/H). Causal ratings and learning scores were subjected to a 2 x 2 Bayesian ANOVA with factors of cue (B/D vs. F/H) and instruction (neutral vs. known cause).

In causal ratings, there was moderate evidence for a main effect of instruction, $BF_{10} = 7576$, *error%* = 0.962, $\eta_p^2 = 0.312$, $F(1,62) = 28.09$, $p < .001$. However, evidence for an effect of cue favoured the null, $BF_{01} = 2.222$, *error%* = 0.967, $\eta_p^2 = 0.031$, $F(1,62) = 2.010$, $p = .161$, and there was evidence to support excluding the interaction between cue and instruction from the model $BF_{Excl} = 3.584$, $\eta_p^2 = 0.004$, $F(1,62) < 1$, $p = .642$. Providing an instruction that one cue in a particular compound is a known allergen appears to have effectively reduced causal ratings for competing cues overall, however, there was no evidence of a blocking effect in causal ratings in either group. Two planned directional Bayesian t-tests were employed to assess blocking (B/D < F/H) in causal ratings within each instruction group. In the neutral group the evidence did not support either hypothesis, $BF_{01} = 1.06$, *error%* < 0.001, $d = 0.263$, $t(31) = 1.49$, $p = .073$ in favour of the null. In the known cause group there was evidence for the absence of a blocking effect, $BF_{01} = 3.085$, *error%* < 0.001, $d = 0.109$, $t(31) = 0.61$, $p = .272$. Overall, this pattern of results differs somewhat from Experiment 3.1, where a robust blocking effect was found in the neutral condition, but is more consistent with the experiments in Chapter 2.

**Figure 3.4**

*Learning scores and causal ratings from Experiment 3.2*




*Note.* Learning scores (top panels) and causal ratings (bottom panels) for the critical cues from Experiment 3.2 as a function of instruction group (neutral vs. known cause). Left panels plot the means and standard errors for the blocked cues (B/D), instructed overshadowed compounds divided into non-instructed (F/H) and instructed (E/G) elements (a distinction relevant only to the known cause group), and the overshadowed control compounds (M-P). Right panels show the distributions of the blocking scores (F/H – B/D) for each

instruction group. Dots represent individual blocking scores, the mean and standard error are given by the horizontal bar and shaded boxes respectively.

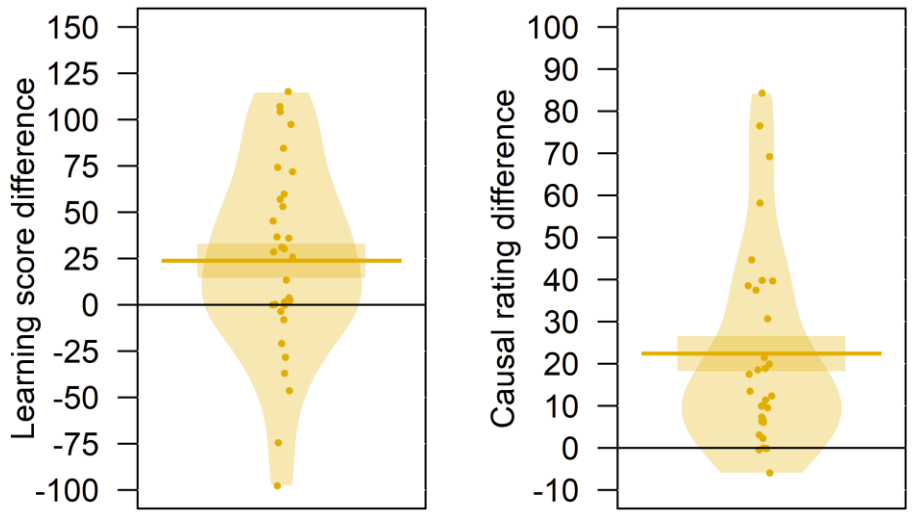
In the learning scores, a complementary Bayesian ANOVA revealed evidence for a main effect of instruction, $BF_{10}$ = 3.813, *error%* = 1.084, $\eta_p^2$ = 0.092, $F(1,62) = 6.25$, $p = .015$, reflecting higher learning scores on average following neutral than known cause instructions. However, the evidence for an effect of cue, $BF_{10}$ = 1.897, *error%* = 0.737, $\eta_p^2$ = 0.081, $F(1,62) = 5.50$, $p = .022$, and of an interaction between cue and instruction, $BF_{Incl}$ = 0.604, $\eta_p^2$ = 0.004, $F(1,62) = 2.52$, $p = .118$, was equivocal. Although the ANOVA did not support an overall blocking effect, planned directional Bayesian t-tests indicated moderate support for the presence of a blocking effect in learning scores in the neutral group, $BF_{10}$ = 8.139, *error%* < 0.001, $d = 0.479$, $t(31) = 2.71$, $p = .005$. Taken together with the apparent absence of a blocking effect in causal ratings, this is consistent with the results from the new patient groups in Chapter 2. The known patient group on the other hand, showed evidence for an absence of blocking in learning scores, $BF_{01}$ = 3.283, *error%* < 001, $d = 0.097$, $t(31) = 0.55$, $p = .293$.

To examine the instructed blocking effect in the known cause group, responses for the elements of the overshadowing instructed compounds that were not designated as allergenic (cues F and H) were compared to those for the overshadowed non-instructed cues (those for which no element of the compound was explicitly labelled as allergenic, cues M to P). The distributions of the difference scores for these comparisons are given in Figure 3.5 (see Figure 3.4 for the means of relevant cues). Two Bayesian paired samples t-tests specifying a directional alternative hypothesis that there would be an instructed blocking effect (M-P > F/H) were conducted over each measure. In both learning scores and causal ratings there was evidence to support the alternative, $BF_{10}$ = 6.644, *error%* < 0.001, $d = 0.461$, $t(31) = 2.61$, $p = .007$; and $BF_{10}$ = 3120, *error%* < 0.001, $d = 0.958$, $t(31) = 5.418$, $p < .001$, respectively. In

other words, both associative memory and causal judgements were weaker for cues competing with a cue explicitly labelled as allergenic than for cues competing with another novel and uninstructed cue.

**Figure 3.5.**

*Instructed blocking in learning scores and causal ratings*




*Note.* Distributions of difference scores comparing the magnitude of instructed blocking (M-P – F/H) in the known cause group on learning scores (left panel) and causal ratings (right panel). Scores larger than zero indicate an instructed blocking effect. The dots represent individual participant's instructed blocking scores, the horizontal bar gives the mean and the shaded box gives the standard error.

## Discussion

Experiment 3.2 provided novel evidence of an instructed blocking effect and successfully generated a pattern of blocking effects in the neutral group that matched what I had observed in the Chapter 2 experiments. However, despite the modifications to the design, Experiment 3.2 did not fully resolve the ambiguity of the results of Experiment 3.1 regarding the blocking effect itself. I will briefly discuss each of these findings in turn.

Consistent with the strong response to known cause instructions during training, explicit instructions about the causal relationships in the task produced something akin to instructed cue competition, both in what was encoded in memory and in the assessment of causality made at test. That is, for cues that were novel when the new patient was introduced, participants showed poorer associative recall if cues were paired with an instructed known cause of allergy than if paired with another novel (and uninstructed) cue. As hypothesised from Experiment 3.1, explicit targeted instructions were able to produce a significant learning bias akin to the bias seen in blocking by pretraining. Participants also attributed the likely cause of the allergy to the known cause in the compound, producing lower causal ratings for the blocked cue than for cues subject to overshadowing without explicit instruction. This may be because they had poorer associative memory for the outcome paired with the blocked cue, as reflected in learning scores, or it may reflect a rational decision that the uninstructed cue was unlikely to (also) have been a cause of the reaction.

The causal ratings provided the clearest pattern regarding the effects of instructions on cue competition in the task. There was evidence that the known cause instructions reduced causal ratings overall. That is, regardless of cue type, participants rated the blocked and overshadowed cues as less likely to be causal if they were given instructions that the competing cue was a known allergen than following a change in patient alone. The analysis further indicated the absence of a blocking effect in general and that this pattern was consistent across instruction groups. In other words, while the known cause instructions produced lower causal ratings for blocked and overshadowed cues in general, neither group showed evidence of blocking in their causal judgements. This pattern differs from that observed in Experiment 3.1, in which the neutral group in particular had a robust blocking effect on causal ratings. In fact, the pattern of results in the neutral group of *this* experiment is ostensibly much more consistent with the results of Chapter 2 than that observed in Experiment 3.1. Recall that in Chapter 2, introducing a new patient after pretraining reduced cue competition in causal judgements but was much less

effective at reducing competition for associative strength relative to the same patient group. Although I did not use a same patient group in the Chapter 3 experiments, the pattern is consistent in that the neutral group displayed poorer retrieval of the outcome paired with the blocked cue than the overshadowed cue whereas their causal ratings for these cues were very similar, indicative of a weak or possibly absent blocking effect.

In the present experiment, learning scores were on average lower when the competing cue (regardless of whether it was novel or pretrained) was identified as a known allergen of the new patient than when the status of the competing cue was not subject to explicit instruction. This was evident both within the known cause group (learning scores for F/H vs M-P) and across groups (group difference in learning scores averaging over B/D and F/H). Although the known patient instructions clearly influenced learning scores—the instructed blocking effect is interesting in its own right—the key research question motivating these experiments was how do explicit instructions about the pretrained cues *interact* with pretraining to affect competition for associative learning? Unfortunately, the analysis did not provide sufficient evidence to answer this question with any confidence. As in Experiment 3.1, the effect on learning scores of pretraining the competing cue (B/D vs F/H) and its interaction with instruction group produced a pattern of inconsistent evidence that evades clear interpretation. Once again, there was clear evidence of blocking in learning scores in the neutral group, and in this instance there was evidence of its absence in learning scores in the known cause group, meaning that in principle it should have been possible to observe a reliable reduction in blocking in the known cause group. However, the Bayesian ANOVA revealed there was insufficient evidence to support the presence or absence of an interaction between the blocking effect and the instruction manipulation, returning a Bayes factor anecdotally in favour of there being no interaction ($BF_{Incl}$ = 0.604). Based on this evidence alone it might be tempting to conclude that causal information delivered by instructions can produce a similar bias in associative memory to pretraining alone, and can influence causal judgements and overt predictions,

but has a less pronounced effect on biases already produced by prior associative history. The reader may note, however, that this is inconsistent with what Figure 3.5 appears to show, and also with the pattern obtained from the follow up t-tests. That is, there was a robust blocking effect in learning scores of the neutral group and an absence of blocking in the known cause group. This pattern is broadly consistent with what was seen in Experiment 3.1 learning scores, where there was an overall blocking effect that was numerically weaker following known cause instructions. As previously noted, there was not enough evidence to support the presence of an interaction in Experiment 3.1 either. These commonalities, and broader conclusions from this chapter will now be further discussed.

## General Discussion

The key aim of this chapter was to examine the interaction between causal information provided by explicit instructions and associative history in the production of associative memory judgements. The experiments herein were intended to replicate the findings of Chapter 2 under conditions in which the instructions themselves provide a clear alternative source of predictions from which to draw, mitigating the general increase in uncertainty that new patient instructions introduce and therefore potentially increasing the participant's motivation to disregard information obtained during prior learning in the pretraining phase. Two experiments pitted the neutral new patient instructions introduced in Chapter 2 against *known cause* instructions that explicitly identified the known allergens of the test patient in an allergist task. While there are several comparisons of interest in these experiments, they were primarily motivated by a hypothesis about the learning scores for cues that compete with the known causes, namely that blocking would be evident under neutral conditions but substantially reduced under the known cause instructions. In this regard, the two experiments produced quite similar results in that both generated reliable blocking effects on learning scores in the neutral group, and both *appeared* to generate weaker blocking in the known cause groups. This was most evident in Experiment 3.2 in which planned comparisons indicated no blocking in the learning

scores of the known cause group. However, in both experiments, the analyses failed to yield a clear answer as to whether blocking was meaningfully reduced by the instruction manipulation, with evidence supportive of neither an interaction nor its absence.

It is clear that the explicit instructions were effective. They successfully influenced learning scores; when allergens were explicitly identified in the instructions, associative memory retrieval was strong for instructed cues and weak for non-instructed cues, with evidence of blocking-like effects based on instruction in both experiments. They also successfully influenced performance (overt predictions made about the new patient's allergies) and causal judgements made at test were generally consistent with the instructed information. However, the question of whether there was some interaction between instructed causal information and associative history remains open.

There *were* some differences in the results of the two experiments. Most notably, the neutral condition of Experiment 3.1 produced a stronger blocking effect in causal ratings than one would expect based on the results from Chapter 2, and seemingly much stronger than Experiment 3.2. Without an appropriate within-experiment comparison, I cannot confirm that the new patient instructions in this group *failed* to modify participants' judgments. Nevertheless, it seems reasonable to speculate why differences in the procedure in this experiment might have altered this result. Experiment 3.1 did not include no allergic reaction trials. This potentially had two unintended consequences, a) that the base-rate of an allergic reaction was higher, and b) that participants may have engaged with the causal learning task in a different way, possibly focusing more on getting their prediction right during training than thinking about the implications of the new patient instructions for the causal status of the cues.

The results do not appear to support the notion that participants rely on the base rate of allergies to judge the causal status of ambiguous cues. As mentioned in Chapter 1, there is some evidence that causal judgements for blocked and overshadowed cues are affected by the base rate of events in the learning task. For instance, Jones et al. (2019) demonstrated that causal ratings for the

blocked cue were influenced by the probability that the patient would suffer an allergic reaction in general, such that higher base rates produced higher causal ratings for ambiguous cues. In Experiment 3.1, the base rate of allergies was 1, that is an allergic reaction occurred after every meal. If participants were relying on base rates to resolve uncertainty about blocked and overshadowed cues, then causal ratings for these cues should be high. Instead, causal ratings for ambiguous cues in the neutral group appear to be comparable to those given when the base rate of allergic reactions was lower due to the addition of the 'no allergic reaction' outcome (in Chapter 2 and Experiment 3.2). Of course, this is a cross-experiment observation so I cannot put any weight behind the comparison but as Figure 3.2 shows, causal ratings for these cues sit around the middle of the scale indicating enduring uncertainty about their causal status despite the high probability of an allergic reaction.

Across both experiments, instructions about known causes led to substantial changes in cue competition, on both learning score and causal rating measures, which I refer to as instructed blocking effects. These effects were relatively strong and observed over a number of comparisons between groups and conditions. These effects show that beliefs about the likelihood of an outcome, established through instruction rather than prior history with the predictive cues, can have a strong effect on memory as well as predictions and judgements. Instructed blocking effects could be a result of changes in stimulus encoding (e.g. stronger attention to known causes at the expense of the competing cue) or changes in the internal predictions that determine the strength of prediction errors (and by extension, the strength of memory encoding). Thorwart and Livesey (2016) referred to influences on learning of this nature as "non-associative" effects to distinguish them from effects of specific learning history (within the context of an experiment, for instance). However, it is noteworthy that they still make use of highly familiar prior information about foods, food categories, and the chance of such foods causing allergies. It might well be the case that in instantiating new predictions using instructions, I am in fact relying on past associations to compete with new ones. I argue that even if this is assumed to be true,

the instructed blocking effects are of interest because they demonstrate the flexibility of instructed cognitive control over our use of prior knowledge, and furthermore demonstrate that this flexible use has implications for the way we encode new associations between other (competing) cues in our environment and meaningful outcomes. Experiments with similar aims in the context of learned predictiveness have found similar effects on learning scores for competing cues, but with clear limitations as instructions often appear to be only partly effective in redirecting learning towards a new set of cues (Don & Livesey, 2015; Shone et al., 2015).

The results from these experiments therefore provide some promising avenues for future research on the use of instructions to influence associative learning. Given the equivocal evidence for interaction between instructed and pretrained predictions, it may seem justifiable to replicate these effects with a much larger sample, as I did for Chapter 2. However, several factors led me to instead move to investigate the issue of instructed control of associative learning in a different way. The poor evidence for interaction in these experiments was central to this decision. Despite there being very weak or no blocking on learning scores in the known cause groups (as predicted), evidence for interaction is hampered by both high variance in the difference scores and limits on the effect size that I am trying to remove with this manipulation (that is, the blocking effect on learning scores under new patient conditions). Chapter 4 therefore explores other ways in which instruction-based control over cue competition in associative learning might be manipulated.

# CHAPTER 4

## Participant goals affect the efficacy of causal instructions

Chapter 2 provided preliminary evidence that at least in one regard, associative memory processes may not be as open to interaction with explicit instructions as I first hypothesised. I found that, in a blocking design, instructed information that successfully reduced the relevance of pretraining (evidenced by the effect on causal judgements and overt predictions), had little impact on learning scores designed to assess associative memory encoding more directly. In other words, if we assume that prediction error is a primary determinant of the strength of associative memory, modifying the relevance of associative history via instructions did not appear to modify the prediction error that drives associative memory processes. However, in Chapter 3, providing explicit instruction about the cues themselves was sufficient to overcome this difference in sensitivity to some degree. I found evidence of an instructed blocking effect, in which telling the participant that the patient was allergic to a specific food resulted in less learning about a competing food. As discussed in the previous chapter, the evidence that instructions *attenuated* a blocking effect achieved through pretraining was less clear. Nevertheless, the more explicit manipulation used in Chapter 3 did appear to be more effective than the new patient instruction used throughout Chapter 2.

There are several reasons why these direct and explicit instructions about causal relationships might be more effective for producing cue competition in learning (one of which is that they are in fact simply taking advantage of previously learned associations). I suggested in Chapter 3 that one reason why these instructions might be more effective is because they provide a competing set of predictions that carry some certainty, rather than increasing the ambiguity of the causal situation, which is effectively what the new patient instructions otherwise achieve. It is possible that prediction error-driven learning is sometimes sensitive to predictions derived by non-associative means, but that achieving this sensitivity requires cognitive effort, which participants may be easily discouraged from engaging in.

Some authors have argued that in tasks such as these, unless given explicit encouragement, participants tend to make intuitive judgements rather than engage in careful reasoning about the causal structure of the task. For instance, as mentioned in Chapter 1, providing instructions that causes are additive in nature seems to increase the magnitude of blocking as these instructions encourage the inferences that support the blocking effect (Mitchell & Lovibond, 2002; Lovibond et al., 2003). However, in the absence of these instructions, or when nonadditive assumptions are encouraged, a reliable and comparable blocking effect endures (Livesey et al., 2019). In other words, without explicit encouragement, participants may fail to take into consideration how causes combine to produce an outcome and instead make a judgement via other means, possibly relying on a heuristic based on memory retrieval. Relying on intuitive causal judgements of this nature might well be considered a rational strategy. In most cases, the strength of an associative memory—which provides an approximation how well one event predicts another—would be informative about cause and effect. Reliance on the strength of such memories (or the ease with which they are retrieved) is a reasonable heuristic for causality that requires little cognitive effort.

Reluctance to engage in complex reasoning in such tasks may reflect a motivational state that is common in the laboratory, where the rewards for doing so are low. This is especially true in tasks that require a prediction and give corrective feedback throughout learning. This format is widely used in predictive learning, including my experiments and others concerned with cue competition in causal learning. Supervised trial-and-error learning tasks of this kind may encourage a performance focused mindset. That is, rather than engaging carefully with the nuances of the causal scenario, participants may focus on the simple goal of accurately predicting the outcome on each trial. These two goals are not mutually exclusive but a focus on getting the answer right in this task may undermine the application of the instructions, if they involve applying a mental model or thinking about the relations between the elements of the task in a resource-intensive way. In other learning contexts, there is evidence that supervised trial-and-error learning leads to poorer discovery and transfer of conceptual rules relative to learning formats that lead to less accurate

predictions (e.g., cf. inference learning in relational categorization; see Goldwater et al., 2018). To the extent that integrating predictions based on instructions and prior information requires similarly effortful processes as those required for relational rule learning, the nature of this supervised learning task may actually *discourage* adjustments to the participant's predictions based on the causal information they are instructed about.

The current chapter examined the hypothesis that the supervised trial-and-error format of standard predictive learning tasks negatively affects the way instructions influence associative predictions and the competitive learning that follows from those predictions. Causal learning tasks instruct participants to engage in a hypothetical scenario, with the assumption that they will construct some form of mental model to assess the causal relationships between those events. However, a participant who is primarily focused on making correct responses may have a diminished motivation (or capacity) for doing so. If the participant's primary goal was to make accurate predictions during training, then instructions designed to alter their assumed model of cause and effect may have little impact on their learning during training. The introduction of a new patient at the training phase was intended to reduce the relevance of pretraining. Disregarding that information, however, would have aided performance for some compounds, especially the blocking compounds that provided some predictive utility despite the change in patient. This may be particularly true for those participants who are intrinsically rewarded for getting a prediction right, they may have little incentive to change the way they interact with the task following the instructed patient change, even if they later take the instruction into account in making causal judgements.

The experiments reported in this chapter tested this hypothesis in two ways. In Experiments 4.1 and 4.2, the instructions regarding the change between pretraining and training stages was manipulated to be framed in a way that was relevant to the participant's task of successful prediction (rather than being exclusively relevant to the causal context, as in previous experiments). In Experiments 4.3 and 4.4, the task-relevance of the instructions was modified indirectly by changing the learning mode from active to passive observation.

## Experiment 4.1

In Experiment 4.1 instructions were used to reframe the pretraining phase of a blocking design as a practice phase. That is, participants were told that the first patient they encountered in an allergist task was just a practice example to familiarise them with the task. To reinforce the manipulation, the instructions emphasised that participants would be tested on their knowledge of the training patient's allergies alone. Describing the pretraining as a practice example (in addition to a distinct learning context as in the new patient groups from Chapter 2) increases the relevance of the instruction to the task of correctly predicting the allergies of the new patient. That is, in the absence of any other information, carrying forward predictions from the pretraining patient is compatible with the goal of getting the predictions right, even if you do not believe it to be relevant to the new context. On the other hand, practice examples tend to have the property of being adjacent to assessed material but distinct enough that the solutions are not applicable to assessed material. It is safe to presume that undergraduate students have had plenty of experience with such procedures, thus there should be less reason to expect that what they have learned in the practice phase is relevant or applicable to getting the predictions right in the next phase (that is, the one that is to be assessed).

Reframing the initial stage of training permitted exploring a second question; is the timing of the instruction (before or after learning occurs) important for its efficacy in influencing learning. Specifically, instructing participants to ignore learned items after the fact (retroactive instructions) may have a different effect to identifying items as irrelevant in the first place (proactive instructions). This is important for two reasons; firstly, the new patient instruction from previous experiments could be considered analogous to directed forgetting instructions if participants assume that what they have learned is irrelevant to what they will learn about a new patient. Studies of list-method directed forgetting, in which participants are instructed to forget a previously studied list of words, have consistently demonstrated the efficacy of such directed forgetting (see Sahakyan et al., 2013 for a review). If the adopted task goal is to accurately predict the outcome on each trial, then

instructing participants that the pretraining phase was just for practice increases the task-relevance of the direction to disregard the pretraining when making predictions in the training phase. Secondly, it is possible that providing proactive instructions to disregard the pretraining phase may change the way in which information is encoded during pretraining; the knowledge that one is completing a practice example may encourage shallow encoding of the associations.

The design was identical to that of Experiment 2.1 with the exception of the instruction groups. The new patient group was retained, and two new instruction groups were introduced, those that received instructions describing pretraining as a practice for the training phase. The practice patient instruction was delivered either immediately before commencing pretraining (proactive practice group) or immediately following the completion of pretraining (retroactive practice group). As in previous experiments, I expected the new patient group to exhibit clear evidence of blocking on the learning scores. The question of interest is whether this blocking effect in learning scores was substantially diminished in the two practice groups. I predicted blocking effects to be absent on causal ratings in all three groups, though the evidence across experiments in Chapter 2 suggest that the new patient instruction may sometimes weaken rather than completely remove the blocking effect on this measure. Therefore, to the extent that practice instructions have a stronger effect on learning scores than the new patient instruction alone, this might be expected to come through on causal ratings as well.

## Method

### Participants

One hundred undergraduate psychology students from the University of Sydney were recruited to participate in exchange for partial course credit. Four participants were excluded for failing to meet the learning criterion of 60% accuracy in each stage of training leaving 96 participants (71 female, *M* age = 19.94) randomly assigned across three instruction groups (n = 32).

**Procedure & Design**

The design and procedure were identical to Experiment 2.1 apart from the placement and content of the critical instructions for the practice groups (for ease, the contingency design is shown again here in Table 4.1). As before, participants were asked to take the role of an allergist diagnosing the food allergies of their patients. They completed two stages of training (pretraining and training) in which they were presented with foods that their patient had eaten and were asked to predict which outcome (nausea, fever, or no allergic reaction) would occur. They were then assessed on their knowledge of the allergies of the patient from the second (training) phase. Foods from training were presented one-by-one, and participants made two judgements about each food. First, they were asked to recall which outcome occurred when the training patient ate the food. Second, they were asked to judge the extent to which the food caused an allergy of any kind in the train patient.

Table 4.1

*Contingency Design of Experiments 2.1 and 4.1*

| | Pretraining | Training | Test |
|---|---|---|---|
| Blocking | A – O1 | AB – O1 | A, B |
| | C – O2 | CD – O2 | C, D |
| Overshadowing | | EF – O1 | E, F |
| | | GH – O2 | G, H |
| Unovershadowing | I – no O | IJ – O1 | I, J |
| | K – no O | KL – O2 | K, L |
| Fillers | | MN – no O | M, N |
| | OP – O1 | U – no O | U |
| | QR – O2 | V – O1 | V |
| | ST – no O | W – O2 | W |
| | | X – no O | X |
| | | YZ – no O | Y, Z |

*Note*. Letters represent cues, randomly assigned to different food items. O1 and O2 represent different allergic reactions ('*Nausea*' or '*Fever*'); 'no O' represents '*no allergic reaction*.'

In addition to the change in patient, the two practice groups were told that the pretraining patient was a practice example and this was emphasised by informing participants that they would *not* be tested on what they had learned during the pretraining phase. A full script of the instructions is provided in Appendix A, but participants in the proactive practice group received the following instructions immediately before commencing pretraining:

> PRACTICE: The first patient you will see is just for practice to give you some experience with the task. You will NOT be tested on what you learn about Mr X's allergies.

To increase retention of this instruction the following reminder was provided after pretraining:

> You have completed the practice trials! Now the real task will begin…

The retroactive practice group on the other hand was informed only after completing pretraining that the pretraining patient was a practice example. The instructions were given as follows:

> You have completed the first part of the experiment. However, the patient you have just seen was a practice example to give you some experience with the task. You will NOT be tested on what you have learned about Mr X's allergies. Now the real task will begin…

**Analysis**

The test data were analysed with a set of planned orthogonal contrasts that assessed the hypotheses. The contrast analyses were conducted using the "BayesFactor" package (Morey & Rouder, 2018) in R (v4.0.3; R Core team, 2020) following the method outlined by Morey (2015). The first contrast assessed whether the framing of the pretraining patient would affect the magnitude of blocking. That is, it compared the magnitude of blocking in the new patient group on the one hand, to blocking in the practice patient instruction groups on average, regardless of when the practice instructions were delivered. The second contrast asked whether the timing of the practice patient instructions, whether they were presented before or after participants had completed the practice training, affected the magnitude of blocking. In other words, the second contrast compared the blocking effect in the proactive practice group to that in the retroactive practice group. For both learning scores and causal ratings, I report the overall blocking effect from a Bayesian ANOVA including all three instruction groups. However, I report the

contrast results for the effect of group and the interaction between group and blocking. The analysis also included planned follow-up tests of the blocking effect within each instruction group. Consistent with previous experiments, I used Bayesian t-tests with an informed prior, the alternative hypothesis was that learning scores or causal ratings would be higher for control cues than for blocked cues.

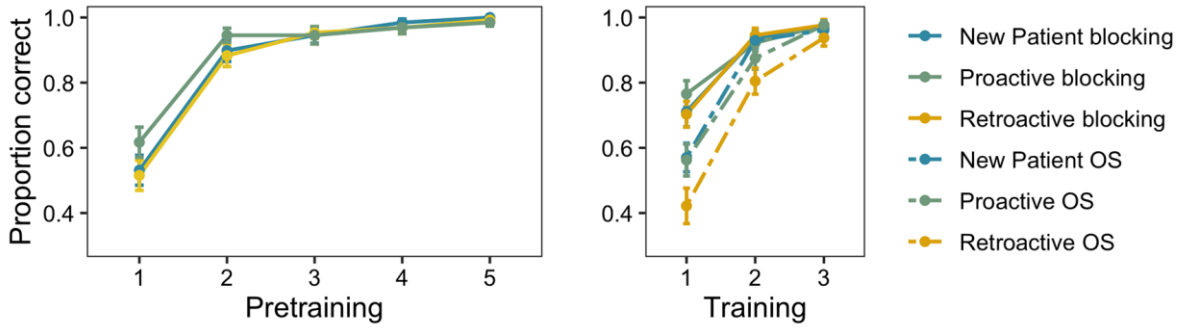
## Results

### Training

All three groups showed evidence of learning the contingencies. In the final block prediction accuracy was above 95% in pretraining, and above 90% in the training phase. Consistent with previous chapters, the performance of the instruction groups in the first block of training after the change of patient was the focus of the training analysis. On average, accuracy was higher for the blocking than the overshadowing compounds, $BF_{10}$ = 490081, *error%* = 0.993, $\eta_p^2$ = 0.234, $F(1, 93)$ = 28.41, $p < .001$. Two orthogonal contrasts on the interaction between cue type and effect of group (comparing accuracy on the blocked vs overshadowed compounds for the new patient vs practice patient, $BF_{Excl}$ = 1.838, $t(93)$ = -1.23, $p$ = .224, and separately for the proactive vs reactive groups, $BF_{Excl}$ = $t(93)$ = -0.82, $p$ = .417) indicated support for the null, suggesting that the observed accuracy benefit for the blocking compounds was present regardless of how the pretraining was described.

**Figure 4.1**

*Accuracy of pretraining and training predictions in Experiment 4.1*




*Note.* Mean (+/- SEM) accuracy across pretraining and training blocks for the blocking and overshadowing (OS) cues and compounds as a function of instruction group (New patient, Proactive practice patient, and Retroactive practice patient).

**Test**

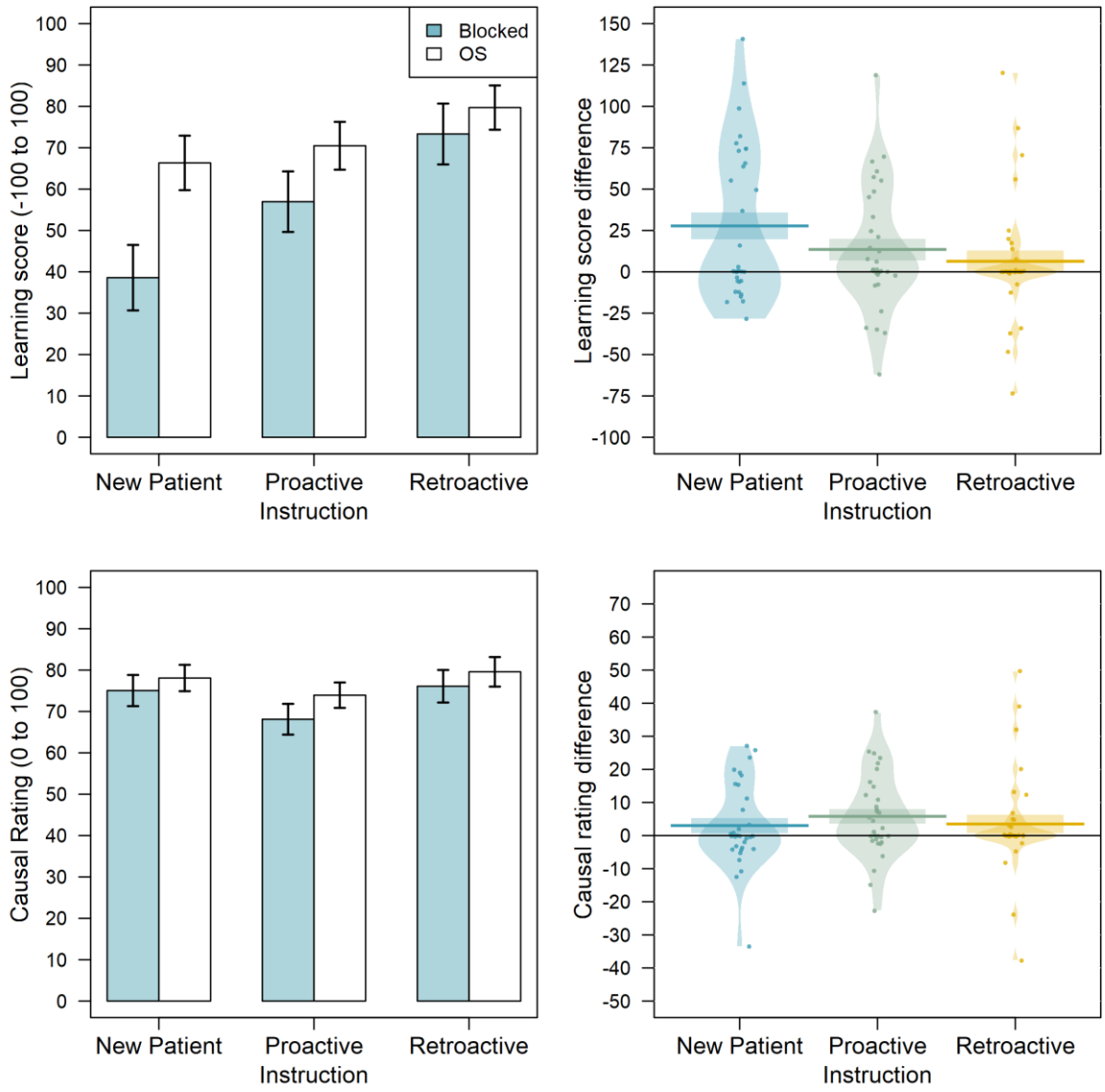
Of greatest interest from the test phase was how the instructions affected cue competition in learning scores (and to some extent, causal ratings). The mean learning scores and causal ratings for the blocked and overshadowed cues are shown in Figure 4.2, alongside the distributions of difference scores for the blocking effect. Difference scores greater than zero indicate evidence of blocking, with higher scores reflecting more complete blocking by the pretrained cue. First, there was strong support for a main effect of cue in learning scores, $BF_{10}$ = 96.871, *error%* = 1.123, $\eta_p^2$ = 0.140, $F(1, 93) = 15.16$, $p < .001$, reflecting evidence of blocking. However, the framing of the pretraining appeared to affect the magnitude of blocking in learning scores, such that blocking was weaker when the pretrain patient was described as a practice example (before or after completing pretraining) than when they were described as simply a different patient, $BF_{Incl}$ = 3.776, $\eta_p^2$ = 0.04, $t(93) = -2.056$, $p = .043$. The placement of the practice instructions before or after pretraining was given did not appear to affect the strength of blocking in learning scores, the data were 3.19 times more likely under the null hypothesis ($\eta_p^2$ = 0.005, $t(93) = -0.715$, $p = .477$). Examining blocking within each instruction group, there was strong evidence for a blocking effect in learning scores in the new

patient group, $BF_{10}$ = 40.35, *error%* < 0.001, $d$ = 0.607, $t(31) = 3.43$, $p < .001$. In contrast, the evidence for blocking in the proactive patient group was weak (albeit significant by conventional NHST standards), $BF_{10}$ = 2.335, *error%* = 0.004, $d$ = 0.364, $t(31) = 2.06$, $p = .024$, and there was weak support for the null in the retroactive practice group, $BF_{01}$ = 2.035, *error%* < 0.001, $d$ = 0.176, $t(31)$ = 0.99, $p = .164$.

Similarly, the Bayesian ANOVA on causal ratings indicated moderate evidence for a blocking effect, $BF_{10}$ = 7.758, *error%* = 2.163, $\eta_p^2 = 0.085$, $F(1, 93) = 8.62$, $p = .004$. The group contrasts revealed that the observed blocking effect in causal ratings was of a comparable magnitude in our instruction groups. Giving new patient instructions rather than practice patient instructions (regardless of when they were given) did not appear to affect the size of blocking in causal ratings, $BF_{Excl}$ = 4.00 in favour of the null, $\eta_p^2 = 0.003$, $t(93) = 0.548$, $p = .585$. Nor did the placement of the practice instructions, there was no difference in the blocking effect in the retroactive and proactive practice groups, $BF_{Excl}$ = 3.37, $\eta_p^2 = 0.004$, $t(93) = -.679$, $p = .499$. Bayesian t-tests on the blocking effect within each group revealed moderate evidence for blocking in the proactive practice patient group, $BF_{10}$ = 6.541, *error%* < 0.001, $d$ = 0.460, $t(31) = 2.60$, $p = .007$. However, there was very weak evidence for the absence of blocking in the remaining groups, $BF_{01}$ = 1.29, *error%* < 0.001, $d$ = 0.238, $t(31) = 1.35$, $p = .094$, and $BF_{01}$ = 1.448, *error%* < 0.001, $d$ = 0.224, $t(31) = 1.27$, $p = .108$, for the new patient group and retroactive practice patient group respectively.

**Figure 4.2**

*Learning scores and causal ratings for Experiment 4.1*




*Note.* Learning scores (top panels) and causal ratings (bottom panels) depicted in two ways. On the left are the mean (+/- SEM) scores for the blocked and overshadowed (OS) cues as a function of instruction group (new patient vs. proactive practice patient vs. retroactive practice patient). On the right are the difference scores (overshadowed – blocked) again as a function of instruction. The points show individual difference scores with the density of scores represented by the beans. Mean and standard error of the blocking scores represented by the central bar and surrounding box respectively.

## Discussion

The training data shows that as early as the first block of the training phase, accuracy was higher for the blocking compounds that contained a pretrained cue than for the novel overshadowing compounds. This indicates that pretraining affected overt predictions during training regardless of how the instructions framed that pretraining. Note however that accuracy for the blocking compounds was affected by the instructions; on average accuracy was numerically higher in the last block of pretraining than the first block of training. This is perhaps unsurprising given the context change and the addition of a novel cue to the blocking contingency at the beginning of training. Nonetheless it suggests some hesitancy to carry forward the predictions from the pretraining phase in all three instruction groups. Ideally, performance for the blocking compounds in these groups would be compared against a group who did not receive a context change. However, in this instance the most we can conclude is that the practice instructions did not increase the pretraining effect in the first block of training relative to a patient shift alone.

There was a very small but reliable blocking effect in causal ratings, and this was present regardless of whether the new patient was described as a practice example or a separate patient. We can contrast this with the pattern in learning scores, where the results indicated that the blocking observed following new patient instructions was reduced when the pretrained patient was described as a practice example. To some extent, it is surprising that *any* of these groups exhibit blocking in causal ratings. The fact that there was a small residual blocking effect across all three experimental groups might suggest participants are reluctant to treat the blocked and overshadowed cues equally in spite of the instructions that should render the pretraining phase irrelevant. Alternatively, it might suggest that the instruction manipulation failed to motivate a change in reasoning for a subset of individuals.

Whether participants knew that they were completing a practice example before or after the fact did not appear to have any impact on how they made judgements at test. This may be because the more task-relevant practice instructions were highly effective at reducing blocking

regardless of when they were given. It may be the case that the two practice instructions have roughly equal efficacy (perhaps one benefitting from a primacy effect, the other a recency effect). In any case, in Experiment 4.2, I chose to use the retroactive instruction, to provide a replication of the result from Experiment 4.1 and to test whether this and the new patient instruction have an additive effect. I will discuss the effect of the practice instruction further in the context of the results of the two experiments.

## Experiment 4.2

Experiment 4.2 was designed to provide a replication of Experiment 4.1 with the addition of a same patient condition to examine how the change in patient context and the practice instruction combine. I used the retroactive instruction condition from Experiment 4.1 only. There were no statistical differences between the proactive and retroactive groups in 4.1, but if anything, the practice instruction appeared to be somewhat more effective when given retrospectively, just before the training phase began. Retroactive instructions also more closely match the timing of the introduction of additional critical information in the new patient condition. Since the causal instruction manipulation (same vs new patient) and the task instruction manipulation (pretraining as practice only) can be introduced independently, this experiment used a 2 x 2 design, with patient instruction (Same vs New patient) and task instruction (practice only vs no additional instruction) as between-subjects factors. Beyond providing a replication of Experiment 4.1, this experiment sought to test whether the patient and practice instructions were additive in their effect on blocking.

### Method

#### Participants

I aimed to recruit 400 participants for this experiment, which was completed online. Participants were recruited from two sources, the pool of undergraduate Psychology students from the University of Sydney (N = 66), and a community sample from the United Kingdom recruited via the online platform Prolific Academic (N = 326). As these experiments were conducted online, additional checks on attention and adherence to the instructions were required. The first of these

was an instruction check; participants were asked to complete a brief survey (see Appendix A) that assessed their understanding of the instructions. Those that failed more than twice were excluded from the analyses. The second was a writing check, given at the end of the experiment to ensure participants correctly followed the instruction to not write anything down during the experiment. Participants were asked to report honestly whether they wrote anything down during the experiment. Those who reported doing so were excluded from the analyses. A total of 71 participants were excluded, seventeen for failing the instruction check more than two times, six who reported writing down the contingencies during training, and 48 who did not reach the learning criterion of 60% accuracy in both pretraining and training phases. All analyses were run on the remaining 321 participants (*M* age = 28.67, 188 female, 119 male, 12 non-binary, 1 other, and 1 who declined to respond). They were randomly assigned to groups yielding *n* = 80 in group same patient, *n* = 83 in new patient, *n* = 76 in same patient + practice, *n* = 82 in new patient + practice.

**Design & Procedure**

The contingency design was the same as Experiment 4.1 (see Table 4.1)[16]. Participants completed the experiment on their own personal computer. They were instructed not to use a phone or tablet and to complete the task in one sitting. They were also instructed to not write anything down during the experiment. Before commencing the experiment, a short survey was issued to confirm that the instructions had been understood. If any question was answered incorrectly, participants were asked to re-read the instructions, were automatically redirected to the first page of instructions, and then given another attempt to answer the survey questions. This continued until they answered all questions correctly.

[16] A small modification was made to the fillers in the design, to bring it in line with Experiment 2.3b in which the baseline of allergic reactions was raised to 0.66. This was a practical decision based on the timing in which these experiments were run and how they were programmed. I will reserve discussion of baseline manipulations to the final chapter of this thesis.

All participants received the critical instructions (same patient only, new patient only, same patient + practice, new patient + practice) after completing pretraining. Following the procedure of Experiment 2.3b, the order in which participants made the memory and causal judgements was counterbalanced. For consistency, I tested for order effects but found evidence supported the null in the case of all interactions and the main effect of test order (smallest $BF_{Excl}$ = 1.94 for the causal instructions x task instructions x test order interaction in learning scores, and $BF_{Excl}$ = 1.733 for the cue x task instructions x order interaction in causal ratings). In all other respects, the experiment procedure was identical to that of Experiment 4.1.

After completing the experiment, they were asked to indicate (honestly) whether or not they had followed the instruction to not write anything down, it was noted that their response would not affect their compensation. Participants recruited via the university pool were sent a completion code to redeem for a small amount of course credit once the experiment ended. Those recruited from Prolific Academic were compensated for their time at rate of £6 per hour. Compensation was issued once recruitment was complete, as per the standard procedure for the Prolific Academic platform. A full script of the instructions can be found in Appendix A.

## Results

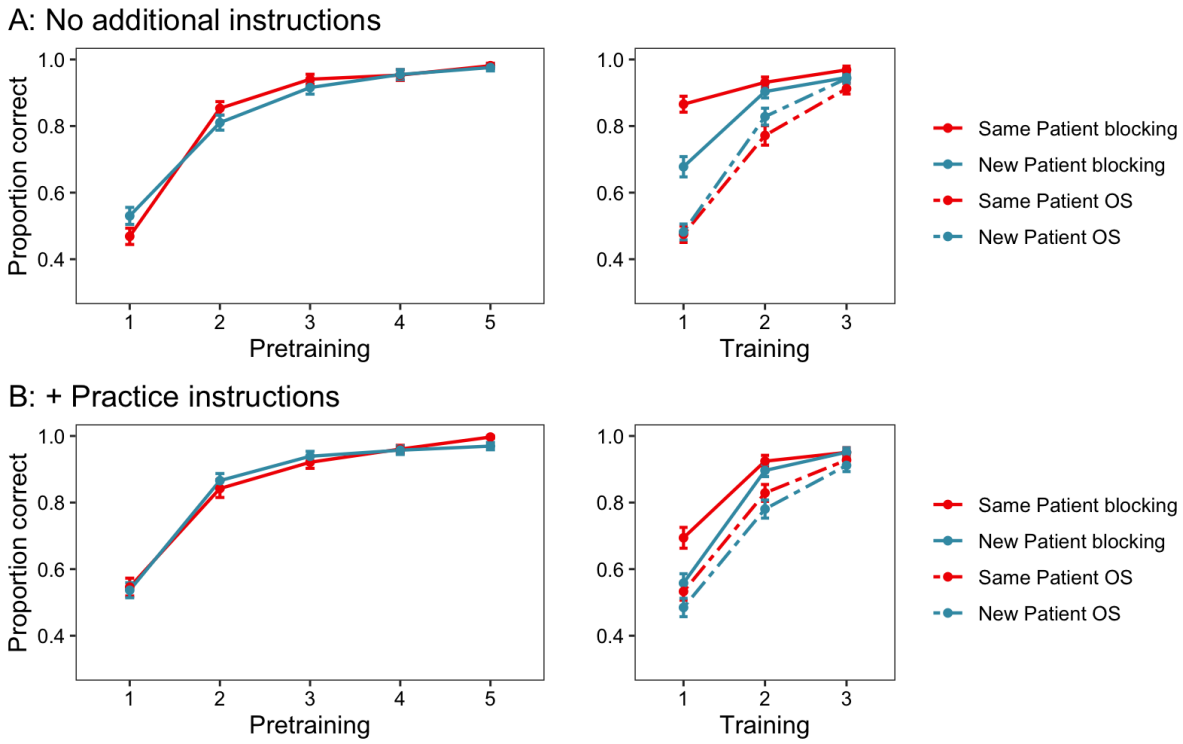
### Training

Participants appeared to learn the contingencies by the end of pretraining and training. As illustrated in Figure 4.3, mean accuracy above 90% and above 80% for all contingencies in the final block of pretraining and training respectively. Training accuracy in the first block of the training phase was subjected to a Bayesian ANOVA with factors of contingency (blocking vs overshadowing), causal instructions (same patient vs. new patient), and task instructions (practice instructions vs no additional instructions). This revealed evidence of a pretraining effect (higher accuracy for contingencies that contained a pretrained element than novel contingencies, $BF_{10}$ = 2.619 x $10^{21}$, *error%* = 1.131, $\eta_p^2$ = .269, $F(1, 317) = 116.81$, $p < .001$). That accuracy benefit appeared to interact with both causal instructions (it was more pronounced when participants were told the training

examples came from the same patient than from a different patient, $BF_{Incl}$ = 50.87, $\eta_p^2$ = .042, $F$(1, 317) = 13.87, $p$ < .001); and separately with task instructions (more pronounced if the pretraining was not described as a practice phase, $BF_{Incl}$ = 7295, $\eta_p^2$ = .063, $F$(1, 317) = 21.49, $p$ < .001). These instruction effects did not appear to interact in the early stages of training, $BF_{Excl}$ = 2.506, $\eta_p^2$ = .006, $F$(1, 317) = 1.98, $p$ = .160 for the cue by causal instruction by task instruction interaction.

**Figure 4.3**

*Accuracy of pretraining and training predictions in Experiment 4.2*




*Note.* The figure shows mean proportion correct predictions during pretraining (left panels) and training (right panels). The contingency is represented by line type, blocking (solid lines) vs overshadowing (OS, dashed lines) and is given as a function of causal instruction (same patient in red vs. new patient in blue). Groups given no additional task instructions are shown in panel A, groups given additional practice patient instructions are shown in panel B. Error bars represent the standard error of the mean.

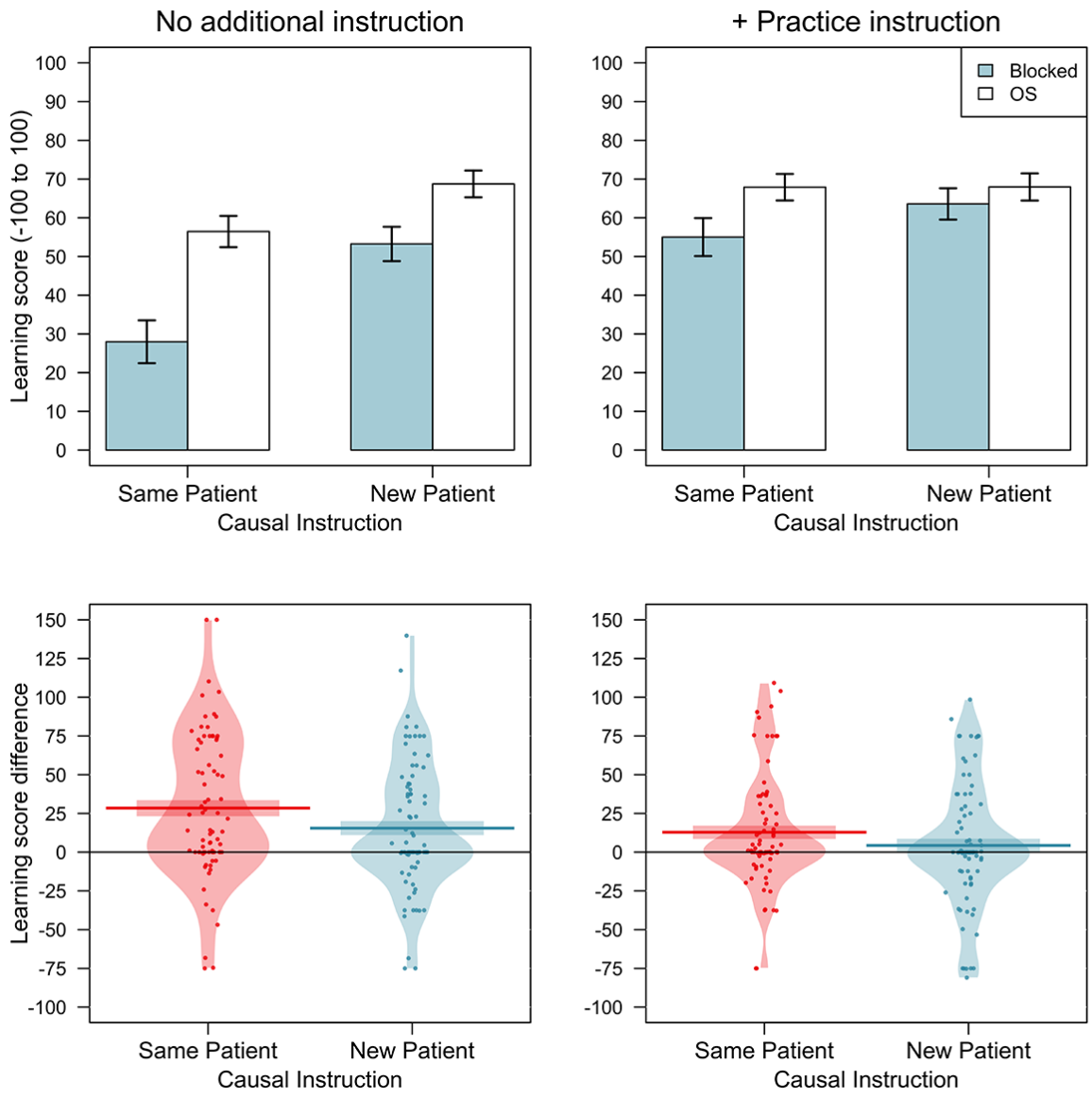
The critical instructions also had separate main effects on training performance with lower accuracy on average given new patient than same patient instructions ($BF_{10}$ = 305, *error%* = 2.308, $\eta_p^2$ = .065, $F(1, 317) = 22.17$, $p < .001$) and given practice instructions rather than no additional instructions ($BF_{10}$ = 4.08, *error%* = 1.086, $\eta_p^2$ = .027, $F(1, 317) = 8.84$, $p = .003$). Consistent with the lack of a three-way interaction, these effects did not appear to interact ($BF_{Excl}$ = 6.25, $\eta_p^2 < .001$, $F(1, 317) = 0.002$, $p = .996$ , for the causal instruction by task instruction interaction).

**Test**

Learning scores were analysed by means of a Bayesian ANOVA with factors of causal instruction (Same patient vs New Patient), task instruction (Practice vs no additional instruction), and cue type (blocked vs overshadowed). Mean learning scores as well as blocking scores for each group are given in Figure 4.4. The ANOVA revealed evidence for a main effect of cue type, with blocked cues receiving lower scores on average than overshadowed cues, $BF_{10} = 3.15 \times 10^7$, *error%* = 0.718, $\eta_p^2$ = .123, $F(1, 317) = 44.56$, $p < .001$. The causal instruction and task instruction manipulations also had main effects on learning scores, with scores lower on average if the pretraining pertained to the same patient ($BF_{10}$ = 19.76, *error%* = 3.96, $\eta_p^2$ = .032, $F(1, 317) = 10.65$, $p = .001$), and higher on average if participants received practice instructions ($BF_{10}$ = 21.45, *error%* = 1.53, $\eta_p^2$ = .035, $F(1, 317) = 11.49$, $p < .001$).

**Figure 4.4**

*Learning scores for Experiment 4.2*




*Note*. Left panels illustrate data from the groups given no additional task instructions, right panels, those given additional practice task instructions. Top panels give the mean (+/- SEM) learning scores for blocked and overshadowed (OS) cues as a function of causal instruction (same patient vs. new patient). Bottom panels show the individual blocking scores (learning scores for the overshadowed cues – blocked cues) and the distribution of those scores (shaded bean). The bolded line represents the mean and the shaded box one standard error around the mean.

The evidence favoured the absence of a three-way interaction suggesting that the two instruction manipulations do not have an interactive effect on learning scores, $BF_{Excl}$ = 5.32, $\eta_p^2$ < .001, $F$(1, 317) = 0.237, $p$ = .627. There was, however, evidence for an interaction between cue and task instructions, such that regardless of whether there was a patient change or not, blocking was overall weakened by describing pretraining as a practice phase, $BF_{Incl}$ = 6.79, $\eta_p^2$ = .026, $F$(1, 317) = 8.48, $p$ = .004. Bayes Factors for the remaining interactions favoured the alternative but the evidence was weak, although note that by conventional NHST standards they are statistically significant ($BF_{Incl}$ = 1.619, $\eta_p^2$ = .017, $F$(1, 317) = 5.49, $p$ = .020 for the causal instruction by cue interaction, and $BF_{Incl}$ = 1.195, $\eta_p^2$ = .013, $F$(1, 317) = 4.178, $p$ = .042 for the causal instruction by task instruction interaction).

Follow-up Bayesian t-tests were conducted to test for evidence of blocking in each of the groups. In line with previous analyses the tests specified a directional alternative hypothesis: learning scores for overshadowed cues would be higher than for blocked cues. There was strong evidence of a blocking effect in learning scores from both the same patient and new patient groups who did *not* receive practice instructions ($BF_{10}$ = 71865, $d$ = - .619, $t$(79) = -5.54, $p$ < .001, and $BF_{10}$ = 54.72, *error%* < .001, $d$ = - .380, $t$(82) = -3.46, $p$ < .001 respectively). However, amongst those given practice instructions, blocking was present when the same patient was used for practice and training but absent for those who saw a patient change for the training phase ($BF_{10}$ = 17.308, *error%* < .001, $d$ = - .349, $t$(75) = -3.05, $p$ = .002, and $BF_{01}$ = 3.064 in favour of the null, *error%* < .001, $d$ = - .110, $t$(81) = -0.992, $p$ = .162 respectively).

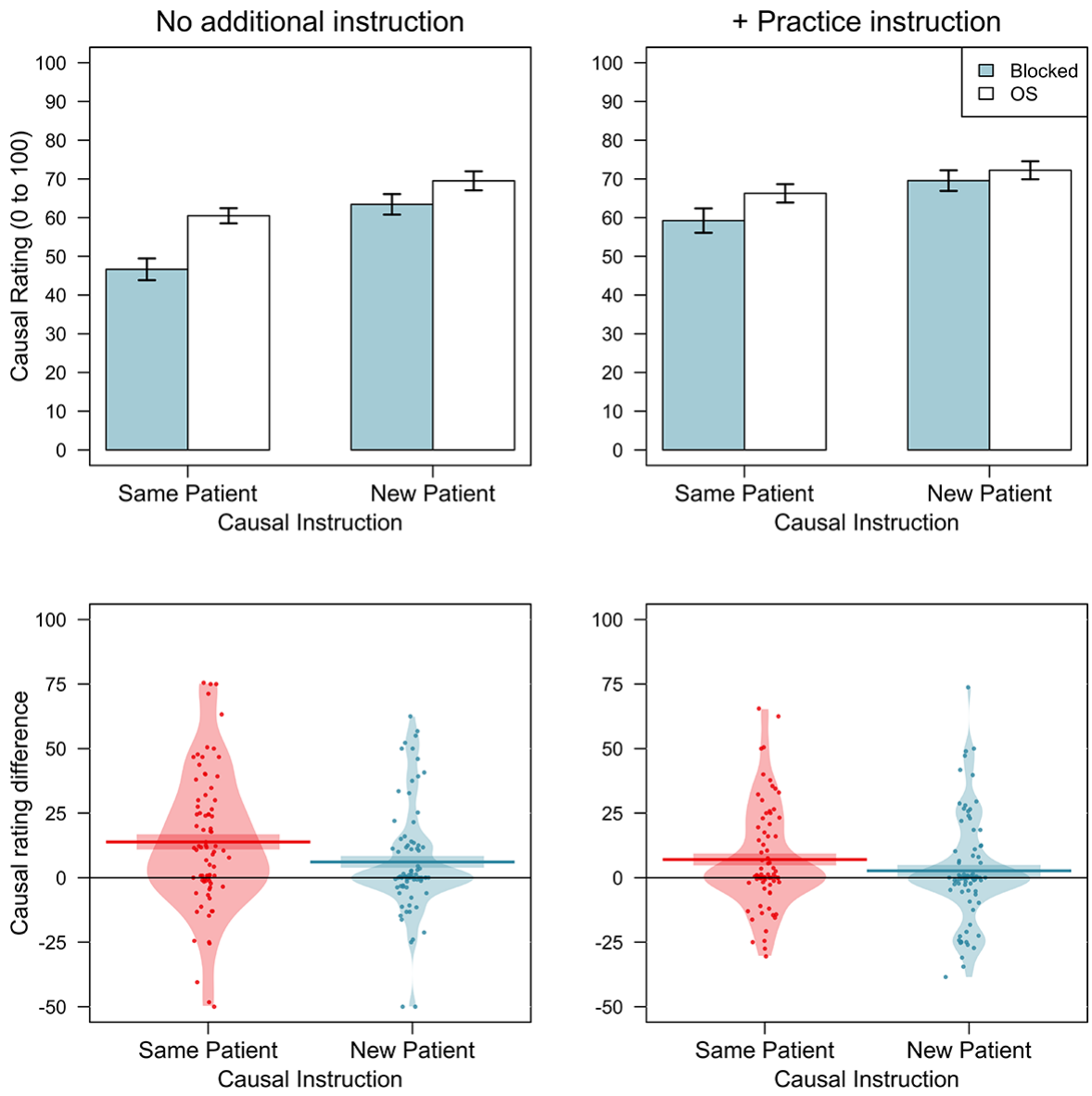
The mean causal ratings and blocking scores as a function of cue type, task instruction and causal instruction are given in Figure 4.5. The same analysis was performed on the causal ratings. Again, there was good evidence for the presence of all main effects. That is, the evidence supported an overall blocking effect in causal ratings, $BF_{10}$ = 1.115 x $10^6$, *error%* = 0.871, $\eta_p^2$ = .103, $F$(1, 317) = 36.52, $p$ < .001. Mean causal ratings were higher on average when the pretraining was described as a practice phase ($BF_{10}$ = 8.542, *error%* = 1.060, $\eta_p^2$ = .028, $F$(1, 317) = 9.14, $p$ = .003), and when

participants experienced a patient change ($BF_{10}$ = 3326, *error%* = 0.814, $\eta_p^2$ = .064, $F$(1 ,317) = 21.85, $p$ < .001). As with the learning scores, evidence supported the null hypothesis that there was no three-way interaction between cue type, cause instructions, and task instructions in the causal ratings, $BF_{Excl}$ = 3.535, $\eta_p^2$ = .001, $F$(1, 317) = 0.48, $p$ = .491. Similarly the null was favoured for the causal instruction by task instruction interaction, $BF_{Excl}$ = 2.972, $\eta_p^2$ = .004, $F$(1, 317) = 1.12, $p$ = .291. However, there was weak evidence of an interaction between cue and causal instruction ($BF_{Incl}$ = 2.358 $\eta_p^2$ = .019, $F$(1, 317) = 6.09, $p$ = .014) reflecting that a patient change weakened the blocking effect in causal ratings regardless of whether the pretraining was described as a practice phase. The evidence did not clearly favour either hypothesis for the cue by task instruction manipulation, $BF_{Excl}$ = 1.135, $\eta_p^2$ = .014, $F$(1, 317) = 4.35, $p$ = .038. However, note that again by NHST standards this would constitute a statistically significant result.

The follow up Bayesian t-tests revealed strong evidence for a blocking effect on causal ratings in the groups that received same patient instructions ($BF_{10}$ = 3194, *error%* < .001, $d$ = - .526, $t$(79) = -4.71, $p$ < .001 for the same patient no additional instructions group, and $BF_{10}$ = 20.58, *error%* < .001, $d$ = - .357, $t$(75) = -3.110, $p$ = .001 for the same patient + practice group). There was also evidence to support the presence of blocking in the new patient group given no additional instructions although the effect size was smaller than in the equivalent same patient group, $BF_{10}$ = 6.595, *error%* < .001, $d$ = - .293, $t$(82) = -2.668, $p$ = .005. In the group given both new patient and practice instructions however, the data were 2.37 times more likely under the null hypothesis, suggesting some evidence for the absence of a blocking effect in this group, *error%* < .001, $d$ = - .131, $t$(81) = -1.19, $p$ = .119. Overall, this pattern is consistent with the learning scores and with the causal and task instructions having an additive effect on blocking, that is new patient and practice instructions both weakening the effect.

**Figure 4.5**

*Causal ratings for Experiment 4.2*




*Note*. Causal ratings as a function of causal instruction (same patient vs new patient) and task instruction (left panels: no additional task instruction; right panels: causal plus practice instruction). Columns represent the mean (+/- SEM) causal ratings for blocked and overshadowed (OS) cues. The bottom panels show causal ratings as a difference score (overshadowed – blocked) but include both individual scores (represented by dots) and the distribution of these scores (represented by the shaded bean). The mean and SEM of these distributions are given by the horizonal line and surrounding shaded box respectively.

## Discussion

Experiment 4.2 found evidence consistent with 4.1 in that adding the task instruction, which described pretraining as a practice phase that would not be assessed at the end, reduced the blocking effect on learning scores. The effect of this instruction on blocking in causal ratings was less clear, with the Bayes analyses not supporting the presence or absence of an effect. The effects of adding a causal instruction to reduce the relevance of pretraining (the new patient instruction) and adding the task instruction did not appear to interact on any measure. This might indicate that each instruction partially increases the probability that the participant will discount prior learning when they are learning (and making judgments) about the ambiguous cues, and that the effects of each instruction in this respect are additive. Evidence of an effect of the causal instruction on blocking was equivocal on both the learning score and causal rating measures when averaging over the task instruction conditions. However, it is worth noting that the effect of the new patient instruction in the *absence* of the practice phase instruction (i.e., the replication of previous experiments in Chapter 2) was, once again, clearer on causal ratings than on learning scores.

Taken together, Experiments 4.1 and 4.2 provide evidence that when the relevance instruction added at the end of pretraining is better aligned with participants' task goals, the instruction has a stronger effect on learning in the subsequent learning phase. This suggests that such instruction *can* affect the competitive learning process that leads to weaker cue-outcome memory for the blocked cue, but only under conditions that motivate the participants to use such information appropriately. This will be discussed in greater depth in the General Discussion. First, I will report an alternative means of interrogating this question explored in Experiments 4.3 and 4.4.

## Experiment 4.3

Experiments 4.1 and 4.2 provided attempts to align the instruction manipulation with the participants' goals by increasing the relevance of the instruction to making accurate predictions. In contrast, and as a complement to this approach, Experiments 4.3 and 4.4 provided attempts to align the participants' goals with the original instruction manipulation by changing the mode of learning.

In these experiments, I eliminated the incentive to get the answer right by removing the requirement to make overt predictions. In Experiment 4.3, I first tested the feasibility of the passive observational procedure, before comparing results from passive and active versions of the task more directly in Experiment 4.4.

Experiment 4.3 employed the same causal instruction manipulation as earlier experiments but employed a passive learning mode. That is, rather than make an active prediction about the outcome of each meal, participants merely observed the outcome of each meal. I predicted that this manipulation would enhance the participant's focus on the causal context of the learning task. I hypothesized that, if successful, this manipulation would result in a more complete removal of the blocking effect (particularly on learning scores) when new patient instructions were introduced.

## Method

### Participants

Sixty-four students (*M* age = 20.86, 40 female) from the undergraduate psychology cohort at the University of Sydney were recruited to participate in exchange for partial course credit. Participants were randomly assigned to either the same patient or new patient groups (n = 32).

### Procedure

During training participants saw each meal that the patient had consumed one by one and were asked to click a button marked 'continue' to reveal the specific outcome that followed that meal. As there was no way to track learning during the training phase, a second test phase was added for those in the new patient group that assessed memory for the allergies of the patient from pretraining. As the same patient group experienced only one patient this additional test was only included for the new patient group. In this test participants were asked to recall which outcome followed each of the cues from pretraining (see Table 4.2 for the design). In all other respects the procedure was identical to that of Experiment 2.1.

**Table 4.2**

*Design of Experiment 4.3*

| | | | Test | |
|---|---|---|---|---|
| | Pretraining | Training | Train Patient | Pretrain Patient |
| Blocking | A – O1 | AB – O1 | A, B | A |
| | C – O2 | CD – O2 | C, D | C |
| Overshadowing | | EF – O1 | E, F | |
| | | GH – O2 | G, H | |
| Unovershadowing | I – no O | IJ – O1 | I, J | I |
| | K – no O | KL – O2 | K, L | K |
| Fillers | MN – no O | U – no O | U | M, N |
| | OP – O1 | V – O1 | V | O, P |
| | QR – O2 | W – O2 | W | Q, R |
| | S – O1 | X – no O | X | S |
| | T - O2 | YZ – no O | Y, Z | T |

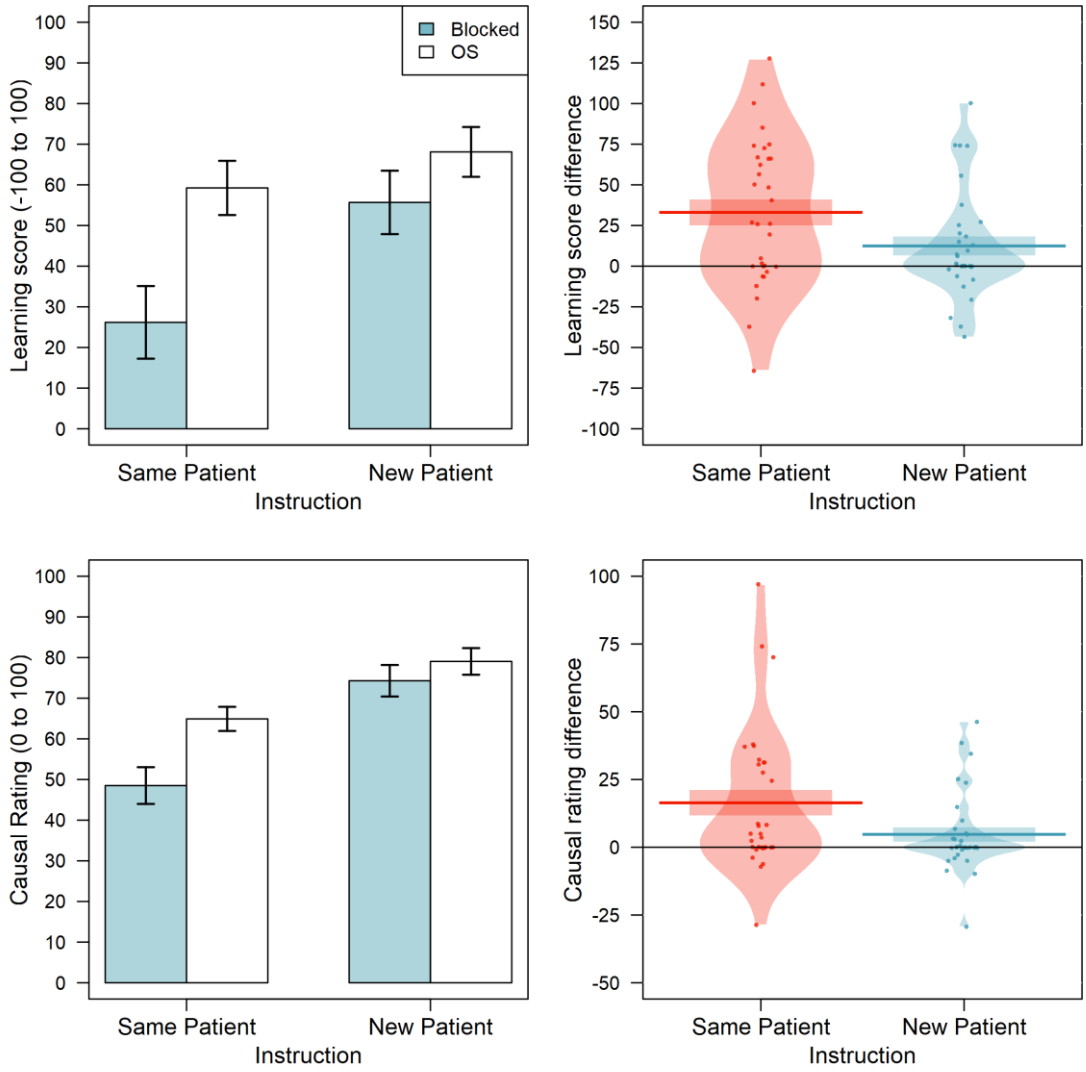
*Note.* Letters represent cues randomly assigned to different food items, O1, O2 and no O represent the different outcomes (two different allergic reactions or *no allergic reaction* respectively). The same patient group did not complete the second test (pretrain patient test) as for them the patient from pretraining was the same as the patient from training.

## Results

Mean learning scores and causal ratings for the critical blocked and overshadowed test cues from the train patient test are given in Figure 4.6. Additionally, all participants from the new patient group (n = 32) completed a recall test for the allergies of the patient from pretraining. The mean of the learning scores for all cues from pretraining was 73.15% (*SE* = 4.86), and the mean score for cues unique to the pretraining patient (cues M – T in Table 4.2) was 77.10% (*SE* = 5.25). There were five participants who scored below 40% on the pretrain patient test, removing them from the analysis did not alter the pattern of results so they were retained in the final analysis.

**Figure 4.6**

*Learning scores and causal ratings from test of training patient's allergies in Experiment 4.3*




*Note.* Learning scores (top panels) and causal ratings (bottom panels) shown as a function of instruction group (same patient vs. new patient). Left panels show the mean (+/- standard error) for the blocked and overshadowed (OS) cues. Right panels show the distributions of the difference scores for the blocking effect (overshadowed – blocked). The mean and standard error of the difference scores is represented by the central line and surrounding box respectively.

**Test phase**

On average, learning scores for the blocked cues were lower than for the overshadowed cues, $BF_{10}$ = 590, *error%* = 1.971, $\eta_p^2$ = 0.255, $F(1, 62) = 21.21$, $p < .001$. The instructions had a main

effect on learning scores, $BF_{10}$ = 1.638, *error%* = 2.524, $\eta_p^2$ = 0.064, $F(1, 62) = 4.23$, $p = .044$, with higher scores on average in the new patient group. However, the model comparison indicated only anecdotal support for an interaction between blocking and the instructions participants received, $BF_{Incl}$ = 1.625, $\eta_p^2$ = 0.066, $F(1, 62) = 4.37$, $p = .041$. Again by conventional NHST standards this would be a significant result suggesting that, there is at least some evidence to suggest that the blocking effect in associative memory was weaker following new patient instructions than following same patient instructions. Nonetheless, this ought to be interpreted with caution given the strength of evidence estimated by the Bayesian analysis. Examining each group individually, there was very strong evidence of blocking in the same patient group, $BF_{10}$ = 233, *error%* < .001, $d$ = 0.734, $t(31) = 4.15$, $p < .001$, and although weaker (as the interaction suggests) there was some evidence to support a blocking effect in the new patient group, $BF_{10}$ = 2.645, *error%* = .004, $d$ = 0.376, $t(31) = 2.17$, $p = .021$.

Averaged across the two groups, there was evidence of a blocking effect in causal ratings, $BF_{10}$ = 74.995, *error%* = 0.954, $\eta_p^2$ = 0.199, $F(1, 62) = 15.42$, $p < .001$. The causal ratings for the blocking contingencies reflected some sensitivity to instructions. Ratings were on average higher in the new patient group, $BF_{10}$ = 510, *error%* = 1.018, $\eta_p^2$ = 0.241, $F(1, 62) = 19.68$, $p < .001$., and there was some evidence (albeit weak) that the magnitude of blocking was modified by the instructions, $BF_{Incl}$ = 1.759, $\eta_p^2$ = 0.070, $F(1, 62) = 4.68$, $p = .034$. Simple effects tests revealed substantial evidence in favour of blocking in the same patient group, $BF_{10}$ = 45.884, *error%* <.001, $d$ = 0.616, $t(31) = 3.49$, $p < .001$, but weak evidence for the alternative in the new patient group, $BF_{10}$ = 1.544, *error%* = 0.002, $d$ = 0.320, $t(31) = 1.81$, $p = .040$. On the whole, it appears that blocking in causal ratings was weakened by the context change introduced by the new patient instructions, though the evidence was not as strong as in some previous experiments.

**Discussion**

In Experiment 4.3, I ran an observational learning version of the canonical design used throughout Chapter 2. In this experiment, when looking at blocking in learning scores, there appeared to be

greater sensitivity to instructions. In fact, this time, the evidence for sensitivity to instruction appeared to be nearly as strong in learning scores as in causal ratings, where a similar effect was evident (though the evidence for interaction was equivocal on both measures). This result stands in contrast to the previous experiments using the patient instruction manipulation, and suggests that removing the proximal goal of making correct predictions may have increased the sensitivity of prediction-error learning to instructions.

Although this result is interesting, Experiment 4.3 by itself provides only tentative evidence, particularly given that it did not provide a direct comparison with active (i.e. supervised trial-and-error) learning conditions. Cross-experiment comparisons are possible given the similarities between the designs but these carry a range of limitations given that the experiments in Chapter 2 were run at a different time and under different conditions. For this reason, and for the sake of replicating the results, I ran a direct comparison between passive and active learning conditions during training in Experiment 4.4.

## Experiment 4.4

The present experiment provided a replication of the results in Experiment 4.3 but extended the design to include both passive and active training modes. This allows for a within-experiment comparison of the training mode manipulation. The design was based off Experiment 4.3, this time crossing the between-subjects factors of learning mode (active vs passive) with patient instruction (same patient vs new patient). Based on the results of Chapter 2 and Experiment 4.3, I expected to find a stronger manipulation of the blocking effect on learning scores under passive learning conditions.

### Method

#### Participants

A community sample was recruited using the online platform Prolific Academic. A total of 367 participants were recruited. I included the same exclusion criteria used for Experiment 4.2 (which was also run online). A total of 44 participants were excluded from Experiment 4.4; twenty-

four for failing an instruction check more than two times, two who reported writing down the contingencies during training, and 18 who did not reach the learning criterion of 60% accuracy in both pretraining and training phases. This left a final sample of 323 (*M* age = 33.10, 186 female, 129 male, 8 non-binary) randomly assigned to groups; n = 83 in the same patient active group, n = 78 in the new patient active group, n = 80 in the same patient passive group, and n = 82 in the new patient passive group.

**Procedure**

Participants were given the same causal scenario in all groups. The experiment followed the same contingency design[17] and procedure as Experiment 4.2 with the exception of the mode of learning. Participants in the active conditions were asked to make trial-by-trial predictions about the outcome based on the cues, as in Experiment 4.2 and earlier experiments. Whereas, participants in the passive conditions were asked to click continue to observe the outcome, as in Experiment 4.3. Participants received either same patient or new patient instructions after pretraining as in the earlier experiments. The order in which participants made the memory and causal judgements was counterbalanced as in Experiments 2.3b and 4.2[18].

## Results

**Training**

Terminal accuracy in the pretraining and training periods for those in the active learning groups was high, indicating that those participants learned the contingencies well. Figure 4.7 shows accuracy as proportion correct predictions for the blocking and overshadowing compounds for those in the active learning groups. As with previous experiments the training period analysis focused on

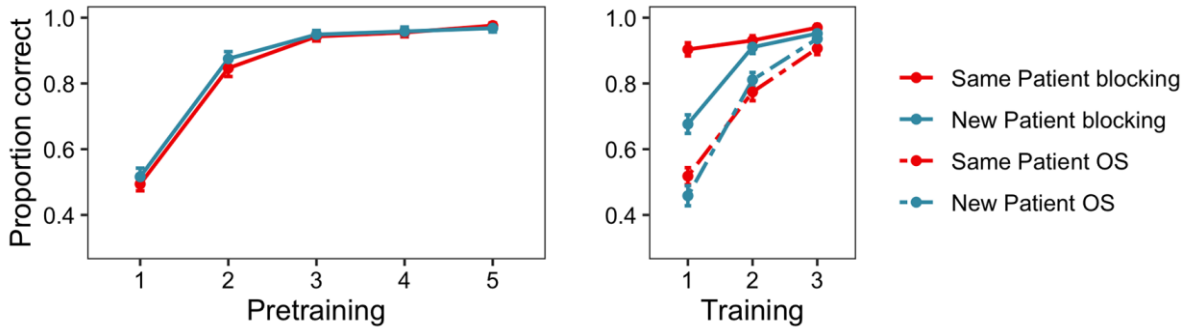
[17] Due to experimenter error the contingency design did not include the additional test of the pretrain patient's allergies given in Experiment 4.3.

[18] To test for order effects in test judgements I ran a Bayesian ANOVA for each measure that included test order as a factor. Although the evidence for excluding two of the relevant interactions was weak ($BF_{Excl}$ = 2.632 for the causal instructions x training format x order interaction in learning scores, and $BF_{Excl}$ = 0.682 for the cue x training format x order interaction in causal ratings), the remainder offered support for the absence of order effects.

the first block of training after the critical instructions were received. A Bayesian ANOVA with factors of contingency (blocking vs overshadowing) and causal instructions (same patient vs new patient was performed over accuracy scores from the active learning mode group. This revealed a significant pretraining effect, $BF_{10}$ = 6.97 x $10^{23}$, *error%* = 1.003, $\eta_p^2$ = .481 , $F(1, 159) = 147.07$, $p < .001$, such that accuracy was on average higher for blocking contingencies after the instructions were given. However, as illustrated in Figure 4.7, there was evidence of an interaction between pretraining and instructions, $BF_{Incl}$ = 47.11, $\eta_p^2$ = .067 , $F(1, 159) = 11.34$, $p < .001$. This reflected a more pronounced pretraining effect when there was continuity between the pretraining and training phases (that is when the same patient, rather than two different patients, was encountered in both phases). The instructions also influenced training accuracy, such that accuracy was poorer on average following the introduction of a new patient, $BF_{10}$ = 651, *error%* = 2.550, $\eta_p^2$ = .147 , $F(1, 159) = 27.44$, $p < .001$.

**Figure 4.7**

*Accuracy of pretraining and training predictions for the active groups from Experiment 4.4*




*Note*. Mean (+/- SEM) proportion correct predictions for participants completing active learning in the pretraining (left panel) and training (right panel) phases. Means are given as a function of contingency (blocking contingencies represented by solid lines, overshadowing (OS) contingencies by dashed lines) and causal instruction (same patient in red, vs. new patient in blue).

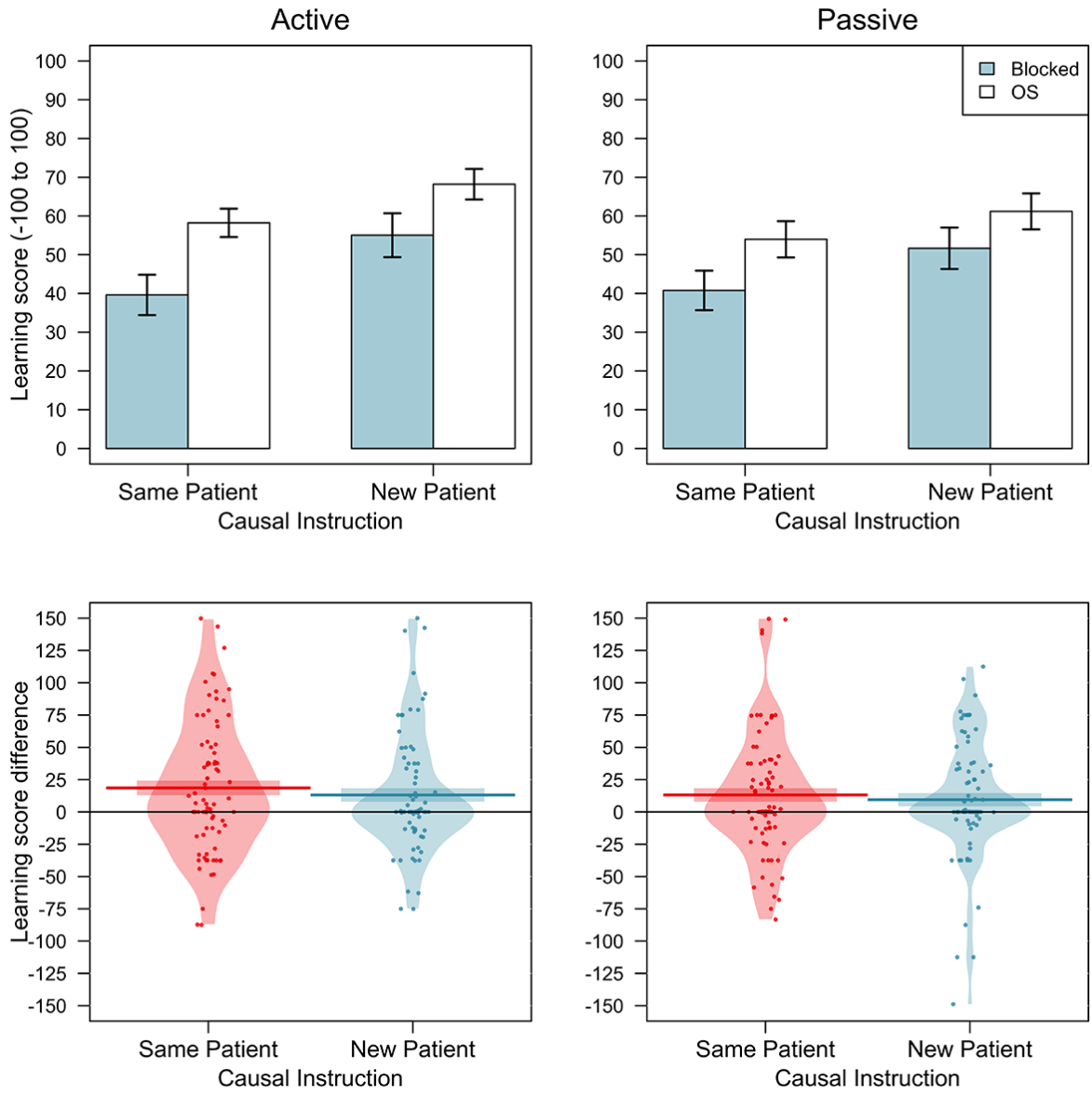
**Test**

Learning scores were subjected to a Bayesian ANOVA with factors of cue (blocking vs. overshadowing), causal instructions (same patient vs. new patient), and learning mode (active vs. passive). Mean learning scores and blocking scores (overshadowed – blocked) are illustrated in Figure 4.8. Consistent with Experiment 4.3 there was support for a main effect of cue, indicating an overall blocking effect in learning scores, $BF_{10}$ = 106741, *error%* = 3.526, $\eta_p^2$ = .077, $F(1, 319) = 26.51$, $p < .001$. In addition, regardless of cue type and learning mode, learning scores were higher on average following new patient instructions than same patient instructions, $BF_{10}$ = 3.859, *error%* = 0.918, $\eta_p^2$ = .022, $F(1, 319) = 7.27$, $p = .007$. All other comparisons favoured the null, the smallest $BF_{Excl}$ = 5.222 for the learning mode by causal instruction interaction, $\eta_p^2 < .001$, $F(1, 319) = 0.204$, $p = .652$.

Turning to the blocking effect in each group, there was evidence of blocking for participants who were engaged in active learning, although the evidence is stronger in the group that did not have a change in patient, $BF_{10}$ = 40.22, *error%* < .001, $d = 0.368$, $t(82) = 3.35$, $p < .001$, for the same patient active group, and $BF_{10}$ = 5.05, *error%* < .001, $d = 0.288$, $t(77) = 2.55$, $p = .006$, for the new patient active group. Among those engaged in a passive learning mode, there was moderate evidence of blocking for the same patient group, $BF_{10}$ = 4.635, *error%* < .001, $d = 0.278$, $t(81) = 2.52$, $p = .007$, but weak support for the alternative following new patient instructions: $BF_{10}$ = 1.191, *error%* < .001, $d = 0.206$, $t(79) = 1.85$, $p = .034$. An independent samples t-test on blocking scores comparing active and passive learning revealed moderate support for the null in both the new patient and same patient conditions, $BF_{01}$ = 5.193, *error%* < .001, $d = 0.080$, $t(156) = 0.50$, $p = .618$, and $BF_{01}$ = 4.713, *error%* < .001, $d = 0.110$, $t(163) = 0.71$, $p = .480$, respectively.

**Figure 4.8**

*Learning scores for Experiment 4.4*



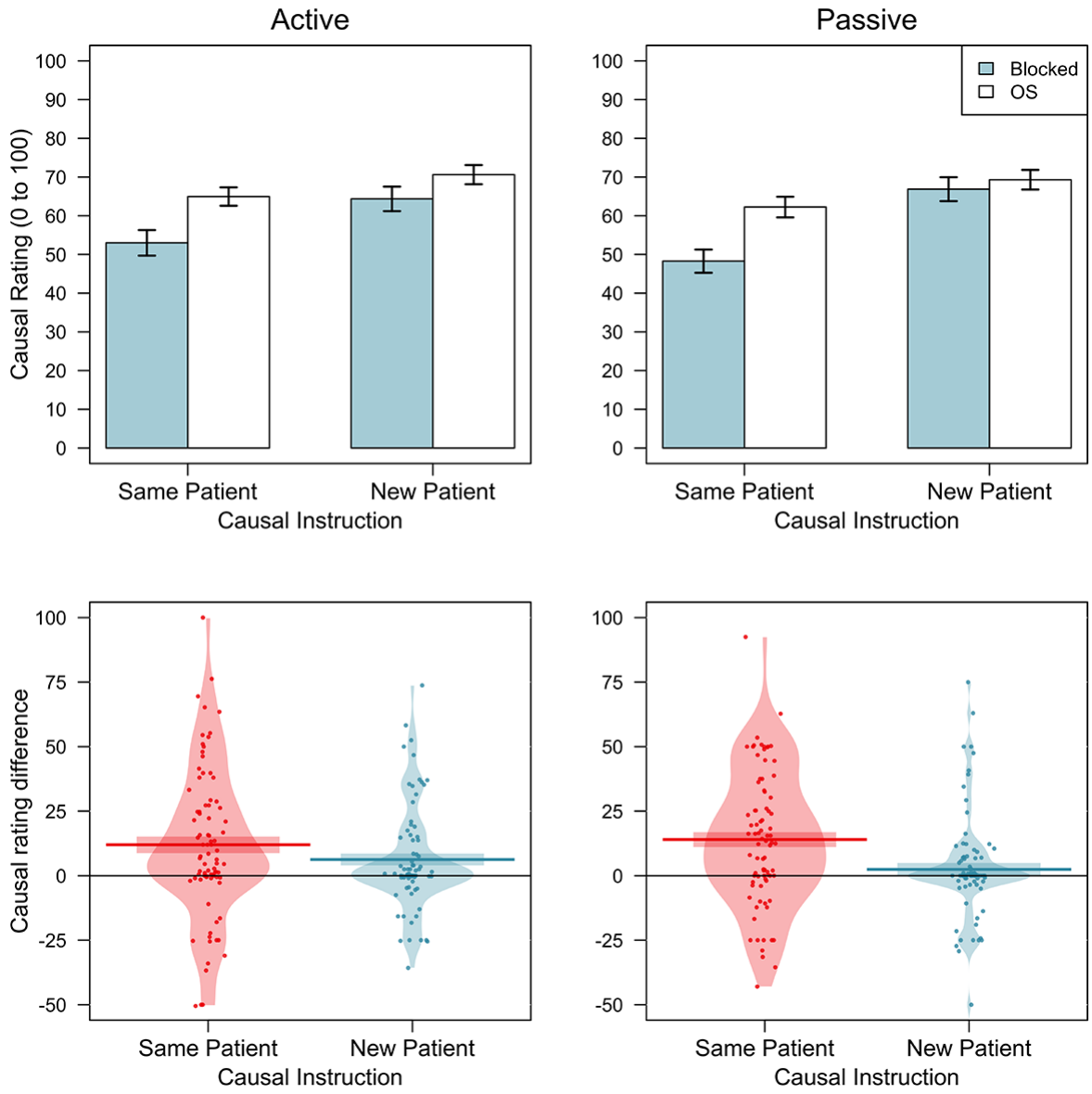

*Note.* Top panels illustrate mean (+/- SEM) learning scores for blocked and overshadowed (OS) cues. Bottom panels show the distribution of individual blocking scores (learning scores for overshadowed – blocked cues). The distribution is represented by the shaded bean, the bolded horizonal line gives the mean and the shaded box surrounding it the standard error of the mean. Learning and blocking scores are given as a function of causal instruction (same patient vs. new patient) and learning mode (active learning in the left panels, passive learning in the right panels).

The pattern observed in causal ratings was largely consistent with that of learning scores with the important exception of the presence of an interaction between cue and causal instruction reflecting stronger blocking in same patient groups regardless of learning mode, $BF_{Incl}$ = 13.85, $\eta_p^2$ = .030, $F(1, 319) = 9.89$, $p = .002$. That is, there was an overall blocking effect in causal ratings ($BF_{10}$ = 7.99 x $10^6$, *error%* = 1.707, $\eta_p^2$ = .111, $F(1, 319) = 39.95$, $p < .001$), and participants gave lower causal ratings on average in the new patient groups ($BF_{10}$ = 709, *error%* = 3.457, $\eta_p^2$ = .054, $F(1, 319) = 18.25$, $p < .001$) but critically and as Figure 4.9 illustrates, the magnitude of blocking was reduced by the new patient instructions regardless of whether participants engaged in active or passive learning of the contingencies. Learning mode had no main effect or interactive effects on causal ratings, smallest $BF_{Excl}$ = 3.484 for the learning mode by causal instruction interaction ($\eta_p^2$ = .002, $F(1, 319) = 0.75$, $p = .387$).

Simple effects tests revealed strong evidence of blocking in same patient groups (same patient active: $BF_{10}$ = 130.87, *error%* < .001, $d = 0.411$, $t(82) = 3.748$, $p < .001$, and same patient passive: $BF_{10}$ = 6325, *error%* < .001, $d = 0.540$, $t(81) = 4.89$, $p < .001$). There was also moderate evidence in support of blocking in the new patient group that engaged in active learning, $BF_{10}$ = 7.729, *error%* < .001, $d = 0.309$, $t(77) = 2.73$, $p = .004$. However, evidence supported the absence of a blocking effect in the new patient passive group, $BF_{01}$ = 3.018 in favour of the null, *error%* < .001, $d$ = 0.111, $t(79) = 1.00$, $p = .161$. Independent samples t-tests comparing the magnitude of blocking in causal ratings following active versus passive learning modes indicated no difference in either causal instruction group, $BF_{01}$ = 5.364, *error%* < .001, $d = 0.73$, $t(163) = 0.47$, $p = .638$ for the same patient group and $BF_{01}$ = 3.223, *error%* < .001, $d = 0.180$, $t(156) = 1.13$, $p = .259$ for the new patient group.

**Figure 4.9**

*Causal ratings for Experiment 4.4*




*Note.* The figure illustrates causal ratings as a function of causal instruction (same patient vs. new patient) and learning mode (active learning in left panels vs. passive learning in right panels). Columns represent the mean and standard error of the causal ratings for the critical test cues (blocked and overshadowed (OS)). The shaded beans give the distribution of individual blocking scores (causal ratings for overshadowed – blocked cues). The mean and standard error of these distributions is given by the horizontal line and surrounding shaded box respectively. Dots represent individual blocking scores.

## Discussion

In Experiment 4.4, I compared the effects of the new patient instructions under active versus passive learning conditions. Experiment 4.3 provided some tentative evidence that observational learning might produce greater sensitivity to the new patient instructions in that blocking in the learning scores appeared to be weaker (in contrast to the results of Chapter 2, where blocking in learning scores were unaffected). However, in Experiment 4.4 it appears that passive and active conditions produced similar results. In particular, the experiment provided another replication of the Chapter 2 results, with reliable blocking and broad *insensitivity* to the new patient instructions in learning scores irrespective of learning format during training. Again, and in stark contrast to the learning scores, causal ratings showed clear sensitivity to the instructions, with blocking dramatically reduced after new patient instructions. If anything, this reduction was slightly clearer under passive learning conditions (blocking in causal ratings was effectively removed in the new patient passive group), but in the absence of interactions with learning mode, I will refrain from interpreting this further.

On the whole, despite the promising (if equivocal) evidence found in Experiment 4.3, it appears that observational learning does not reliably improve the learner's sensitivity to instructions when it comes to how they learn about competing cues and their associated outcomes. The efficacy of this manipulation, and general conclusion from this chapter, will be further considered in the General Discussion.

## General Discussion

In this chapter, I explored a particular hypothesis potentially explaining why blocking of associative learning might be insensitive to the new patient instruction (as shown in Chapter 2) despite the fact that associative memory appears to be affected by explicit instructions in other contexts (e.g. Chapter 3). I hypothesized that cognitive effort might be required to apply instruction-based predictions to modulate the way learned associations are stored and retrieved, and that people might be reluctant to apply that effort if the instruction increased the ambiguity of the reasoning task and was not fully aligned with the participant's current goal of making accurate overt

predictions (i.e., getting the answer right during the training phase). I explored two methods of better aligning the instruction with the proximal goals of the learner. In Experiments 4.1 and 4.2, the instruction was augmented by emphasizing that the pretraining phase was just a practice phase that would not be assessed. The intention was to make the instruction relevant to the way the participant approached the learning task, assuming that they may be more motivated by making correct choices rather than engaging in effortful reasoning about cause and effect. In Experiments 4.3 and 4.4, I removed the need to make overt predictions with participants instead engaged in a passive observational learning task. The intention was to remove the proximal goal of getting the answer right in the hope that it would shift focus towards reasoning about cause and effect.

In all of the experiments in this chapter, the key test of sensitivity to instructions was a reduction in the blocking effect in learning scores. In Chapter 2, blocking in learning scores was robust and resistant to instructional modulation. Therefore, this is the key measure on which I was interested in enhancing sensitivity to the instructions. In contrast, blocking in causal ratings consistently showed evidence of some sensitivity to the new patient instructions. Therefore, in Experiments 4.1-4.4, in all conditions where new patient instructions are given (which includes control conditions in these experiments), the expectation was that blocking in causal ratings would be substantially reduced. This potentially limits my ability to see further modulation of blocking on this measure due to a floor effect.

**Effect of passive learning**

Experiment 4.3 provided some tentative but promising signs that participants undergoing an observational learning procedure could integrate causal instructions in a way that produced an effect on associative memory and not just causal ratings. Although the evidence was anecdotal according to Bayesian analysis, it appeared that blocking of learning scores was substantially reduced in the new patient condition compared to the same patient condition. Unfortunately, when I attempted to replicate this effect and provide a direct comparison with active learning in Experiment 4.4, the evidence suggested that passive and active learning methods did not differ.

Rather, blocking in learning scores was robust and was not affected by learning mode or causal instruction, yielding results very similar to those in Chapter 2.

By removing the need to make active predictions my intent was to encourage participants to engage critically with the causal instructions in the task. However, it must be acknowledged that reducing active participation in the learning task likely has a negative effect on task engagement. One possible explanation then for the lack of clarity around the effect of learning mode on associative memory encoding is that the benefits of passive learning might be outweighed by the costs to task engagement. Although the results reported here do not lend themselves to any clear conclusions, this provides a preliminary exploration of such effects. It remains to be seen whether delivering information in this format, or perhaps in other ways, can affect instructional control of associative memory encoding.

**Effect of practice instructions**

Of the two manipulations examined in this chapter, it would have to be said that the practice instruction had a stronger and more consistent effect on blocking in learning scores. In both Experiments 4.1 and 4.2, the addition of the practice instruction led to a reduction in blocking. Experiment 4.1 suggests that the timing of this instruction is not critical (proactive and retroactive version had an equivalent effect on blocking) and Experiment 4.2 suggests that, to the extent that new patient instructions had *any* effect on learning scores, the effects of cause-focused and task-focused instructions might be additive as there was no indication of interaction between them.

Although clearly effective, these instructions are also specific to the context of performing a learning task in the laboratory. The goal of making correct choices (and receiving regular correct feedback) is arguably not generalizable to situations regularly encountered in everyday life. As such, one might question the relevance of this effect to understanding causal learning and causal reasoning. However, I argue they are relevant for two reasons.

First, they reveal something more general about the parameters under which instructions can affect simple learning and memory mechanisms. It appears that understanding the inferences

gained from those instructions (and being able to apply those inferences to causal judgements) is not sufficient to ensure that they have an effect on memory. Rather the influence that they have on memory may rely on a combination of their utility for the learner's current goals and/or an additional degree of cognitive effort.

Second, although the task itself is artificial, there are many situations in everyday life where generating accurate beliefs about cause and effect is a goal that competes with other more proximal tasks. For instance, many dietary supplements occupy contentious space as forms of complementary and alternative medicine, with limited evidence supporting their efficacy. But at the point where the consumer is exposed to such products or chooses to purchase them, they may not be prioritizing thoughts about their effectiveness over other proximal goals (e.g. observing social etiquette with someone who advocates for their use, or simply finishing the grocery shopping as quickly as possible). Consider any situation in which our current goals compete with the long-term objective of forming accurate memories and beliefs. If we have trouble discounting certain experiences that *should* be considered irrelevant to forming memories and opinions, then this is surely of broad relevance. Moreover, an understanding of how to mitigate this insensitivity could be useful.

**Conclusion**

In this chapter, I attempted to enhance the effect of a relevance instruction on blocking by better aligning the instruction and the participants' current goals. One of these manipulations—aligning the instructions to the participants' current goal of making correct predictions by framing the relevance of prior learning in terms of practice versus the "real" task—appears to have been successful in reducing the blocking effect observed in learning scores. I interpret this effect as suggesting that prediction error-driven learning can be influenced by predictions derived by non-associative processes but, for this influence to occur, it is not sufficient for the individual to simply be exposed to conflicting causal information. Rather, modifying predictions in a way that is meaningful for learning may require the effortful application of cognitive control, which the individual may only engage in if the instruction aligns well with their current objectives. This is a

speculative interpretation, and future research will be required to test several of its assumptions, as well as its generalisability to explicit associative learning more broadly.

# CHAPTER 5

## General Discussion

This thesis was primarily concerned with understanding the processes involved in human causal learning. Chapter 1 laid out the evidence for three key assumptions about human causal learning that underpin the specific aims of this project. First, that there is sufficient evidence to assume that both reasoning and associative memory processes are implicated in causal learning. Second, causal judgements, are sometimes based on aspects of associative memory retrieval, whether that be the strength of the memory linking two events or the fluency with which one event retrieves another. Third, that prediction error largely determines the strength of the association between two events. Importantly, the studies in this thesis were not intended as a test of these assumptions. Rather, the assumptions informed a specific question about how memory and inferential processes might interact to produce causal learning. The central aim of this thesis was thus to take manipulations that should be expected to influence inferential reasoning and test the extent to which they affect both causal judgments and associative memory. Assuming that there is a partial correspondence between causal learning and associative learning processes, there are numerous ways in which these processes may interact (see Thorwart & Livesey, 2016, for a discussion). The focus of this thesis was on the potential for other information to modify predictions and thereby modify the encoding of associative memories, specifically where this information comprises explicit knowledge gathered from reasoned inference or any other source that is not from direct experience with contingencies. If the prediction error learning mechanism itself operates over such predictions in the same way as it does over those derived from direct experience with contingencies, changing predictions in this way should affect prediction error and thereby the strength of associative retrieval. The following chapter will summarise the findings and discuss the results in the context of associative learning theory.

Chapter 2 established a procedure to achieve the central aim of this thesis, namely the modification of prediction error by explicit information. The cue competition design, specifically the

blocking effect, was selected as a test bed for modifying prediction error with instructed information. As a reminder, blocking occurs when learning an association between a novel cue (B) and an outcome (+) is attenuated if the novel cue is trained in compound (AB+) with a cue previously established as a reliable predictor of the outcome (A+). The assumption of some associative models is that the B+ association is poorly encoded because the prediction error is minimal on AB+ trials due to the pretraining of A (Rescorla & Wagner, 1972). This structure provides the opportunity to examine the potential impact of reasoning on associative memory encoding (and on causal judgements that are sometimes based on the output of these processes). Specifically, the aim of Chapter 2 was to shift the predictions that drive prediction error learning by introducing instructions that modify the relevance of the pretraining before blocking occurs. To that end, critical instructions were introduced to a blocking design that was embedded in the Allergist task. In the typical procedure, participants learn about the food allergies of their patient (Miss Y) by observing the pattern of allergic reactions experienced after eating various foods. In Chapter 2, the typical procedure was compared to one in which the instructions introduced a change in patient after pretraining such that the pretraining of A is rendered less relevant to the compound training (AB+). Reducing the relevance of pretraining in this way should in theory increase prediction error on AB+ trials thereby reducing the blocking effect if the predictions that determine associative memory encoding are sensitive to instructed information about the learning context (that is, if they are subject to controlled cognition in the same way that causal judgements appear to be).

All three experiments in Chapter 2 employed this design with minor variations. What emerged was that although the instructions successfully changed participants' causal judgements and (in two of the three experiments) the overt predictions they made during the training phase, they did not appear to affect the strength of associative memory. That is there was a robust blocking effect in learning scores (a measure of associative memory retrieval) regardless of causal instructions. I argued in Chapter 2 that this suggests that prediction error learning is less susceptible to control by instructions than causal judgements.

Chapter 2 involved a “new patient” instruction manipulation in which prior knowledge about associative relationships were rendered irrelevant (or at least substantially *less* relevant) to new learning by instructing the participant that they were observations from a different patient. While this manipulation was effective in changing causal judgements, it essentially involves placing the participant in a state of increased uncertainty while they make predictions during training. Given that uncertainty about predictive relationships tends to be aversive (e.g., participants prefer to learn about outcomes that have been predictable in the past rather than unpredictable outcomes, Hartanto et al., 2021), participants may not be motivated to disregard previously learned information during new learning simply because the task instructions indicate that it is logically irrelevant. In Chapter 3, I thus examined an additional instructional manipulation, which took the form of a very explicit instruction about the allergic reactions suffered by the new patient, with the intention that these instructions would replace strong beliefs established by prior learning with a new set of instructed causal beliefs (which were consistent with prior learning in some circumstances and not in others). In these experiments, I observed clear evidence of a blocking-like effect produced by instruction alone; when the participant was instructed that one food in a compound caused the patient’s allergic reaction, their learning scores for the other cue were reduced. Evidence for interactions between prior learning and instruction were equivocal, however. This method of manipulating predictions is a promising avenue for future work, however the procedure is also difficult to interpret for a number of reasons. In particular, to the extent that causal beliefs established with instructions do affect associative retrieval, it is unclear whether this is because there is flexible cognitive control at play or because the instructions activate previously learned associations (about common food allergies, for instance) very effectively. I ultimately decided to try a different approach for the final series of experiments.

Chapter 4 explored the possibility that the supervised trial-and-error procedure that is typically employed in causal learning experiments of this nature prevents relevant instructions from impacting associative memories. These experiments were motivated by the hypothesis that being

required to make predictions and receiving explicit feedback about their accuracy generates a competing goal of simply performing well upon the prediction task, which might compete with participants updating their mental model of cause and effect appropriately when they encounter relevant causal information in a different (instructed) format. In this chapter, I ran two experiments investigating the effect of manipulating the relevance of the pretraining phase in a way that was aligned with this competing goal, namely by describing the pretraining phase as a practice phase that would not be tested. These experiments provided some evidence that the instruction could indeed reduce blocking in learning scores. In a further two experiments, I also explored the addition of the new patient instructions under passive learning conditions, where the competing aim of making accurate predictions should presumably be removed (or greatly reduced if the participant continued to make trial-by-trial predictions covertly). This manipulation yielded equivocal evidence across two experiments. There was some evidence of an effect of causal instruction on blocking of learning scores in Experiment 4.3, in which I ran the passive condition only. However, this effect was not replicated when I ran a larger study directly comparing passive and active learning in Experiment 4.4.

In the following sections, I will summarise several overarching observations from these experiments and attempt to make sense of them. I will also consider several theoretical issues that may be considered relevant to the interpretation of the results.

**Learning scores display relative insensitivity to relevant instructions**

Across the experiments, it was generally evident that the causal instruction manipulations had a stronger and more consistent effect on explicit judgments of cause and effect (causal ratings) than they did on explicit judgments of cue-outcome association (learning scores). This result was surprising and ran counter to the initial hypothesis that predictions used to guide learning would incorporate information from predictive cues *and* other relevant sources of information.

There is a relatively straightforward interpretation of these results if one takes my initial assumptions that causal judgements are partly determined by the strength of associative learning

but that the two might be independent in some respects. In this case, memory might serve as a basis for the inferences that support causal judgements, but this relationship is largely unidirectional. That is, relevant memories and causal instructions both provide information that is used to draw causal inferences, but those inferences are not necessarily represented by the cognitive machinery responsible for remembering cue-outcome relationships and thus the information conveyed in instructions has a relatively small impact on the strength of associations retrieved from memory.

This interpretation follows from the assumptions that I described at the outset of this thesis. Naturally, these assumptions could be partly or wholly incorrect, and it stands to reason that other interpretations of the data are possible. Previous theorists have, at times, attempted to provide a unified account in which all causal inferences are based on a single associative learning process, or alternatively a unified account in which all associative learning is based on inferences about explicit beliefs (e.g., De Houwer, 2009; Mitchell et al., 2009). These theoretic approaches might, in principle, offer an explanation for the results. As I noted, the intention of this thesis was not to provide evidence for or against such models. However, if the pattern of responses observed consistently over this series of experiments can be shown to be generalisable and replicable in other contexts then they could be viewed as a test bed for model development.

As with any theory, the challenge in reconciling these results with a single-process model would be to provide a coherent account of the conditions under which different types of responses would be expected to be sensitive to instructions and prior learning history. Here I have interpreted these different sensitivities as being a result of the informational content inherently tied to each type of response (e.g. associative memory questions do not invite explicit reasoning processes in the same way as causal judgements). Certainly, there is evidence that the formulation of the test question affects sensitivity to manipulations of causal structure. For instance, one study found that the instructed causal model only affected cue competition when the test question asked about causality, not when it asked about predictive or diagnostic validity (Matute et al., 1996). Similarly, the mode of the test judgement, whether it is given as a discrete choice or as a rating on a

continuous scale, appears to influence the extent to which judgements reflect the influence of non-associative information (Don & Livesey, 2018). Alternatively, there may be differences in the sensitivity of the measures based on their other properties that could result in causal context instructions affecting blocking more on causal ratings than in learning scores. It does not appear that the different sensitivities to instructions are due to the order in which the ratings were presented; although not all of my experiments counterbalanced the order of the two questions, counterbalancing had no effect in any of the experiments that did. However, there are other differences between the measures that might affect their sensitivity.

Future research could explore these remaining possibilities further by systematically varying the mode of each judgment, asking causal and memory-based questions in different formats. This could provide some indication whether there are inherent properties of the rating scales and binary predictions that affects how blocking presents itself on these measures, and how sensitive each is to causal instructions. For similar reasons, and to test the generality of the effects observed here, it would also be useful to examine a range of manipulations designed to vary uncertainty and ambiguity in different ways (for instance using probabilistic relationships between cues and outcomes, differences in the number of learning episodes, or alternative instruction-based manipulations).

This is not the only circumstance in which overt judgements do not align with apparent patterns of memory retrieval or conditioned responding. A well cited example can be found in the Perruchet effect, in which a clear dissociation emerges between reasoned (if irrational) overt expectancies and conditioned responding (McAndrew et al., 2012; Tran, et al., 2020; Weidemann et al., 2016). This thesis represents an attempt to explore this distinction in a novel way, one in which the learning and decision processes involved are clearly explicit rather than implicit. The procedure tested herein provides a new means of examining the relative independence of overt judgements of causality and the strength of associative memory retrieval.

**Associative predictions, prediction error and cognitive control**

This work was motivated by the idea that the strength of associative learning is proportional to prediction error, and prediction errors might be based on more than just the strength with which one event brings to mind another (that is, associative retrieval). My results suggest this is not always the case. Instructions that affect overt predictions and causal judgements do not always affect the strength of cue-outcome learning and retrieval. While it might be tempting to conclude that this means associative learning and inferences based on instructions are completely independent, the results of Chapters 3 and 4 provide some evidence that learning *is* affected by instructed changes in beliefs under some circumstances. I suggested in Chapter 2 that participants may not be highly motivated to "give up" their prior learning when presented with new cue-outcome contingencies in the training phase because doing so provides even less basis on which to make predictions. The implication is that incorporating inferences or instructed changes in belief into learning might require effort and, thus, motivation.

Theories of associative learning often assume that associative change is a continuous process, one which is engaged when we experience co-occurring events of any form (e.g. McLaren et al., 2014). In contrast, it is reasonable to assume that inferential reasoning is engaged on an as-needed basis, depending on the learning context, the individual's goals, and current capacities. With this in mind, it is worth considering the extent to which causal learning tasks motivate participants' inferential reasoning. The properties of typical causal learning experiments may influence how participants engage with the task, encouraging or discouraging careful consideration of the causal structure that is presented. For instance, when studying cue competition effects, the causal relationships in the task are often deterministic (those in which the cause produces the effect every time) rather than probabilistic. This may facilitate the fast, online updating of causal beliefs rather than drawing focus to correctly predicting the effect given the cause. Combining this property with causal ambiguity in the form of blocked and overshadowed cues may serve to encourage inferential reasoning, especially at test when one is asked about the causal status of redundant or causally

ambiguous cues. There is both a high degree of certainty about the relationships that can be observed (e.g., the compound of the pretrained and blocked cue *always* lead to the same outcome) as well as unresolvable ambiguity in the precise nature of this relationship because the two cues are always presented together. These two properties (deterministic causal relationships and causal ambiguity) were present in all experiments reported herein, and thus they should have increased the opportunity to observe the impact of instructed information on prediction error.

In everyday life, causal relationships are rarely experienced in such a controlled fashion. Effects are often probabilistic and fluctuate in severity and timing relative to their possible causes. At the same time, the need to engage in inferential reasoning (and the appropriate way to do so) may not always be as evident as it is in a learning experiment in which people volunteer to participate. Therefore, the conditions in which these experiments have been run may artificially err on the side of encouraging careful reasoning about the causal status of the cues. With that in mind, it is noteworthy that even under the current conditions, when it came to blocking in learning scores, evidence of an impact of instructions was strikingly difficult to find. In Chapter 2, overt predictions and judgements of causality appeared to update with new (non-associative) information, but the strength of associative memory did not. Evidence for an influence of instructions was more forthcoming in some of the experiments presented across Chapters 3 and 4, which used the new patient condition as a starting point for further manipulations, but even here the evidence was equivocal at times. Perhaps tellingly, the clearest evidence came from experiments that framed the change instructions in terms of a practice phase that was not relevant to the learning goals of the participant. Experiments 4.1 and 4.2, which used these instructions, found more convincing evidence of a reduction in blocking on learning scores, compared to the other manipulations that I tried throughout this work.

This raises an interesting question: why would aligning the instruction with the goal of correctly predicting the outcome be particularly effective? One interpretation is that engaging in inferential reasoning is effortful such that even when the task properties encourage it, participants

may be reluctant to apply the required effort, particularly if doing so is at odds with their current goal. Consistent with this idea is evidence that participants can require significant encouragement to apply inferential logic to a causal learning task. For instance, Lovibond et al. (2003) note that in a pilot study of the seminal additivity training manipulation, participants given additivity instructions failed to apply the logic and so demonstrated a similar blocking effect to those given nonadditive instructions. It was only when causal additivity was explicitly demonstrated in a pretraining phase that participants successfully applied the inferences to their judgements. Of course, it cannot be known if this was due to failure to take in the instructions or to a lack of motivation to apply them. However, the latter interpretation is consistent with the finding that non-additive pretraining is insufficient to remove the blocking effect (Lovibond et al., 2003; Livesey et al., 2019). Indeed, López et al. (2005) found that causal model effects, in which reversing the typical causal structure of predictive tasks produces a reduction in blocking, were only present when the importance of causal structure information was particularly stressed to the participants.

If it is the case that causal beliefs update quickly and dynamically but their influence on learning processes requires effort, then future research could manipulate motivation with performance incentives. For instance, if this hypothesis is correct then incentivising participants to understand the causal structure versus performing well on the prediction task should result in a stronger influence of instructions (e.g. the new patient manipulation) on associative memory, as indexed by learning scores. For similar reasons, one might also expect individual differences in intrinsic motivation to understand concepts (versus simply getting the correct answer) to be associated with greater sensitivity of learning scores to instructional manipulations (e.g. see Goldwater et al., 2018). It may be the case that motivation to apply cognitive control in tasks which can be solved (at least in terms of getting predictions right if not resolving causal ambiguity) without the application of logical reasoning is something that differs among individuals. Livesey et al. (2013) found that blocking under nonadditive pretraining was negatively related to performance on the Cognitive Reflection Task (CRT), which is a short questionnaire that captures the tendency to make

intuitive versus more considered judgements (Frederick, 2005). In other words, those inclined to think carefully about the causal structure of the task were less likely to show blocking when causes were demonstrably non-additive. Future research could employ measures like the CRT and the Need for Cognition scale (Cacioppo & Petty, 1982) to identify individuals that are more or less motivated to engage in critical reflection or analytical thinking and therefore may be more or less likely to apply the effort needed to exercise cognitive control over learning.

The effect of time is another consideration that was not explored in this thesis. The experiments reported herein all employed an immediate test of learning. These conditions are arguably optimal for any judgment made at test that requires remembering precisely when (or in which context) events were experienced. For instance, after a delay, a participant that had been given blocking contingencies combined with the new patient instructions may have greater difficulty recalling that the pretrained cue was observed in a different patient to the blocked cue. For this reason, causal ratings might well be less sensitive to the type of instructions used in these experiments if the test is delayed. One could also assume that adding a delay between training (in which the critical instructions are embedded) and test would decrease the sensitivity of causal judgements to instructed information in general, potentially bringing them more in line with the patterns of responding observed in learning scores. Thus, there are reasons to predict that the effect of new patient instructions on causal judgements might fade with a delay. Employing a delay of an hour or a week, for instance, between training and test could provide insight into the longevity of the effect of new patient instructions on causal learning. If it were the case that causal judgements come to resemble the strength of associative retrieval after a delay, then this would have important implications for how these judgements are made outside of the laboratory setting. For instance, intuitive and even more considered judgements may reflect the extent to which all relevant information is recalled at the point when the causal rating is made.

On the other hand, if the effects observed in Chapter 2 are preserved after a delay, then it may suggest that causal beliefs—influenced, as they are, by the new patient instructions—are

updated and stored in a long-term fashion at the time of learning. In other words, it is possible that causal reasoning (i.e., updating of beliefs about potential causes) occurs entirely “on the fly”, during training, and thus isn’t strongly affected by a delay before test ratings, even though the causal ratings and learning scores consistently show different effects. Again, this would provide insight into the processes driving the formation of causal judgements during learning, and how well those judgements are retained in memory. Whether the learner is forming causal beliefs about the blocked cue online during training or at the point of judgement, it might be possible for these terminal causal beliefs to be stored in long-term memory even if the specifics of the information that informed them (the change in patient) is forgotten.

If it is true that competition for associative memory does not align with competition in our explicit causal beliefs, it raises questions about how our experiences in daily life might continue to impact our decisions and judgements even when better or more complete information is available. The statistical landscape of our environment is enduringly complex and making sense of our experience is not a trivial task. Intuitive judgements and heuristics in many cases are useful but are infamously error prone (Tversky & Kahneman, 1974, ; Stanovich & West, 1998). The results of Chapter 4 suggest that one way to improve memory updating with relevant information might be to align the information more closely with the learner’s current goal (e.g., getting predictions right in a fairly arbitrary learning task). In everyday life, this is not likely to be practicable in many situations and the question then arises whether an individual needs to be particularly motivated to exert cognitive control over the learning process in order for prior (and perhaps now irrelevant) information to be ignored. If one assumes that our decisions and judgements about the causal structure of our environments are made on the basis of associative memories, and that in the absence of proper motivation those memories are informed by associative history alone, then one implication is that these memories may reflect inaccurate information and may be difficult to modify. Finding ways to improve the use of relevant information while not being distracted by irrelevant experiences is, of course, a major challenge for cognitive psychology broadly.

**Attention, prediction error, and cognitive control**

The account of the findings of this thesis that I have given thus far has described the effect of instructions on associative memory encoding as resulting from changes in associative strength proportional to a summed prediction error. That is, my interpretations have focused on the learned changes to the processing of the outcome on each trial. However, models of attention that instead describe learned changes to the extent to which the cues are processed also provide a coherent account of blocking (and cue competition effects more broadly) that has received some empirical support. These models assume that prediction error produces learned changes in selective attention to cues, that constitute or contribute to cue competition. The model proposed by Mackintosh (1975) for instance explains blocking as an increase in attention to the pretrained cue – the predictive cue that reliably signals an important outcome – during pretraining and consequent decrease in attention to the blocked cue – the less predictive (redundant) cue – during training. These changes in attention influence the rate of learning such that the blocked cue-outcome relationship is poorly acquired.

It is likely that attentional shifts like those described by Mackintosh (1975) are at least partly responsible for cue competition effects, including blocking. As discussed in earlier chapters, blocking is accompanied by changes in overt attention. Participants spend less time looking at the blocked cue on average than the pretrained or control cues (Beesley & Le Pelley, 2011; Kruschke et al., 2005; Wills et al., 2007). The decrement in attention to the blocked cue also retards later learning when the blocked cue is presented as a valid predictor suggesting blocking is associated with a failure to process the blocked cue (Beesley & Le Pelley, 2011, Kruschke & Blair, 2000; Le Pelley, et al., 2007). There is strong evidence that changes in selective attention play a role in other competitive learning biases such as the learned predictiveness effect (Le Pelley et al., 2016) and the inverse base-rate effect (Don et al., 2021).

However, it should be noted that, to the extent that cue competition involves changes in attention, it is unlikely that these changes are wholly responsible for the learning deficits observed.

Some early models of attention attribute cue competition effects entirely to competition for attention rather than to changes in associative strength resulting from prediction error. In Mackintosh's (1975) conception, cue competition does not require a summed error term. Instead, prediction error produces changes in the associability of each cue separately by means of an individual error term. Such changes in associability affect the attention paid to the cues and consequently the rate at which they are learned about. Although Mackintosh's model of attention has received empirical support, the model in its original formulation cannot account for key phenomena like conditioned inhibition. More successful attention-based models assume that cue competition happens at the point at which an association is formed, driven by both changes in associative strength and changes in associability that are governed by a summed error term (the EXIT model, Kruschke, 2001, 2003; and the hybrid models proposed by Le Pelley, 2004, and Pearce & Mackintosh, 2010).

Experiment 2.2 (Chapter 2) included a measure of overt attention; participants completed the allergist task with the new patient instruction manipulation while their eye movements were recorded. Consistent with past research, patterns of overt attention reflected an attentional bias away from the blocked cue. Critically, as the learning scores reflect, the instructions did not affect this bias. However, perhaps unsurprisingly there was some indication of a small effect of new patient instructions in the early stages of training, consistent with the patterns of overt predictions that participants made during the task. The difficulty in interpreting this data lies in the reciprocal relationship between learning and attention, the direction of causation cannot be known. Does the attentional bias affect the rate of learning or do the patterns of attention reflect what is currently being learned? Hybrid models like those described above posit that both processes, attentional shifts and prediction error learning, are contributing to associative memory encoding. On the basis of Experiment 2.2 we might conclude that both are resistant to instructed control. That is, manipulating the relevance of associative history by instructions does not appear to affect the updating of attention.

While the lack of group effects on learning scores observed in Chapter 2 suggest resistance to instruction in both attention and association formation, the effects of instruction on learning scores that I *did* observe in Chapter 3 and 4 could be attributed to there being some flexibility in selective attention, or the predictions that control association formation, or both. For instance, in Experiments 3.1 and 3.2, perhaps participants focus their attention strongly on known causes when this information is not competing with learning from the pretraining phase, and thus these experiments found evidence of an "instructed blocking" effect. At the same time these experiments yielded only equivocal evidence of an effect of instructed causes on the blocked cue itself, and this might reflect weaker control of attention when learning history provides information that the learner can continue to draw upon. A similar potential explanation could be offered for the results of Experiments 4.1 and 4.2, where describing pretraining as a practice phase appeared to strengthen the effect of the instructions on learning scores. In this case, it is possible that participants motivated to learn the correct answer may apply cognitive control to "balance out" attention between the pretrained cue and the blocked cue, effectively learning about the two cues as if the pretraining had not been experienced.

These effects may, therefore, represent increased sensitivity to instruction in one of at least two of cognitive processes. Where prediction error learning is successfully controlled by instructions, the question remains how such control is achieved, and attention is a clear candidate. Nevertheless, evidence from a variety of experimental paradigms suggests that learned attention biases are not easily overcome by instructions alone (Cobos et al., 2018; Don & Livesey, 2015; Le Pelley et al., 2013; Le Pelley et al., 2015; Shone et al., 2015). For this reason, it would be useful for future work to investigate attention changes further, in the context of the instructional manipulations used in Chapters 3 and 4. Concomitant changes in the distribution of gaze and the manipulation of blocking in learning scores would be of particular interest. Additionally, it may prove informative to examine the timescale of these changes in attention. Previous work suggests that a blocking procedure changes the extent to which the cues capture attention (Luque et al. 2018), and this kind of rapid

attentional capture is likely not easily brought under cognitive control (Cobos et al., 2018; Le Pelley et al., 2015). If, for instance, the findings of Chapters 3 and 4 reflect the influence of controlled attentional processes, to what extent do early or rapid biases persist? Measures of gaze duration are not necessarily equipped to identify these rapid shifts, and thus future work might benefit from the inclusion of a measure of attentional capture.

**Uncertainty and the blocking effect**

Jones et al., (2019) note that the blocking design generates a level of uncertainty about the causal status of the blocked cue. In the absence of the conditions necessary to *deduce* that the blocked cue is non-causal (Livesey et al., 2019), there is a certain level of causal ambiguity about the blocked cue that cannot be resolved. This is equally true of the overshadowing control cues against which the blocked cue is typically compared, but on the basis of conditional probabilities, there is a greater likelihood that either of the overshadowed cues is a cause than the blocked cue. Livesey et al. (2013) noted that the difference in these two conditional probabilities is inversely proportional to the likelihood that any given cue causes the outcome (the base rate of the outcome). Testing the sensitivity of causal ratings to differences in these probabilities, Jones et al. (2019) varied the base rate of the outcome, using filler cues to manipulate the probability that any given cue would be causal. They found that probability ratings for blocked cue approximated the overall base rate of allergic reactions, suggesting that uncertainty about the causal status of the blocked cue is a contributing factor to whether participants display a blocking effect. In an extension of the role that uncertainty might play in the way we evaluate potential causes, researchers have recently proposed the *theory protection* principle to account for the way people appear to update their beliefs when presented with new contingency information (Spicer et al., 2020; 2022). The theory protection principle proposes that participants will attempt to protect beliefs they have previously developed about the causal status of a cue if faced with new information that potentially conflicts with that theory. This type of process is naturally related to the updating one would expect when an individual experiences prediction error. In the case of blocking, however, new information about the

compound does not conflict with the beliefs that the participant would normally generate about the pretrained cue and thus there may be little motivation to engage in theory protection, *per se*.

Regarding the experiments I have reported here, changes in uncertainty presumably accompany the "new patient" instructions, though they would only affect the blocked cue indirectly. The change in patient should affect uncertainty about the likelihood of the outcome in general as well as the specific likelihoods that the individual has estimated for the pretrained cues. Again, the noteworthy result in relation to uncertainty is that while causal ratings seem to be consistently sensitive to this instruction, learning scores were not. To the extent that changes in participants' causal ratings of the blocked cue reflect changes in their level of uncertainty, the manner in which they recalled which cue was associated with which outcome appeared to be substantially less affected. To date, no work has examined the effect of likelihood manipulations (like the base rate manipulation employed by Jones et al., 2019) on learning scores measured independently of judgments of cause and effect. To the extent that these processes are independent it might be informative to investigate whether they are subject to the same biases.

**Conclusion**

The aim of this thesis was to investigate the potential for an interaction between prediction error learning and other (non-associative) information. To that end, I provided instructions that manipulated the relevance of associative (or predictive) history to current predictions. My focus on the interaction *between* associative and reasoning processes in causal learning, rather than on providing evidence for or against the involvement of one or the other process, was spurred by the call to move beyond the debate between single-process models of causal learning and those that maintain a role for associative memory processes (see for example Le Pelley et al. 2017; Thorwart & Livesey, 2016). If one assumes that both processes are involved in determining the trajectory of causal learning, more work is needed investigating the boundary constraints of each process. What roles do they play and what are the limits of their influence on decisions, memory, and judgment? The work in this thesis provides consistent evidence that 1) the blocking effect in causal judgments is

robust but sensitive to instructions about the relevance of past learning history, and 2) the blocking effect in learning (as indexed by judgements of cue-outcome association) is not as strongly affected by these instructions. Furthermore, I found preliminary evidence that motivation to use relevance instructions might be important for whether they influence prediction error learning. By providing some insight into whether and under what circumstances associative memory and causal reasoning processes interact in learning, I hope that this work contributes to our understanding of causal cognition and its relationship with associative learning.

**REFERENCES**

Aitken, M. R., Larkin, M. J., & Dickinson, A. (2000). Super-learning of causal judgements. *The Quarterly Journal of Experimental Psychology Section B*, *53*(1), 59-81. https://doi.org/10.1080/713932716

Aitken, M. R., Larkin, M. J., & Dickinson, A. (2001). Re-examination of the role of within-compound associations in the retrospective revaluation of causal judgements. *The Quarterly Journal of Experimental Psychology Section B*, *54*(1b), 27-51. https://doi.org/10.1080/713932745

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147-149. https://doi.org/10.3758/BF03334492

Allan, L. G. (2003). Assessing power PC. *Animal Learning & Behavior, 31*, 192-204. https://doi.org/10.3758/BF03195982

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation, 14*(4), 381-405. https://doi.org/10.1016/0023-9690(83)90024-3

Arcediano, F., Matute, H., Escobar, M., & Miller, R. R. (2005). Competition between antecedent and between subsequent stimuli in causal judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 228 –237. https://doi.org/10.1037/0278-7393.31.2.228

Barrett, L. C., & Livesey, E. J. (2010). Dissociations between expectancy and performance in simple and two-choice reaction-time tasks: A test of associative and nonassociative explanations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 864–877. https://doi.org/10.1037/a0019403

Beckers, T., De Houwer, J., Pineño, 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*(2), 238–249. https://doi.org/10.1037/0278-7393.31.2.238

Beesley, T., & Le Pelley, M. E. (2011). The influence of blocking on overt attention and associability in human learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(1), 220-229. https://doi.org/10.1037/a0019526

Blanco, F., Baeyens, F., & Beckers, T. (2014). Blocking in human causal learning is affected by outcome assumptions manipulated through causal structure. *Learning & behavior, 42*, 185-199. https://doi.org/10.3758/s13420-014-0137-y

Cacioppo J. T., Petty R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116-131. https://doi.org/10.1037/0022-3514.42.1.116

Carr, A. F. (1974). Latent inhibition and overshadowing in conditioned emotional response conditioning with rats. *Journal of Comparative and Physiological Psychology, 86*(4), 718–723. https://doi.org/10.1037/h0036159

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological review, 104*(2), 367-405. https://doi.org/10.1037/0033-295X.104.2.367

Cheng, P. W., & Holyoak, K. J. (1995). Complex adaptive systems as intuitive statisticians: Causality, contingency, and prediction. In H. L. Roitblat & J-A Meyer (Eds.), *Comparative approaches to cognitive science* (pp. 271-302). MIT Press.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99*(2), 365-382. https://doi.org/10.1037/0033-295x.99.2.365

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Cobos, P. L., Vadillo, M. A., Luque, D., & Le Pelley, M. E. (2018). Learned predictiveness acquired through experience prevails over the influence of conflicting verbal instructions in rapid selective attention. *PloS one, 13*(9), e0200051. https://doi.org/10.1371/journal.pone.0200051

Corlett, P.R., Mollick, J.A. & Kober, H. (2022). Meta-analysis of human prediction error for incentives, perception, cognition, and action. *Neuropsychopharmacology. 47*, 1339–1349. https://doi.org/10.1038/s41386-021-01264-3

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1-20. https://doi.org/10.3758/LB.37.1.1

De Houwer, J., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(3), 965-985. https://doi.org/10.1080/02724980143000578

De Loof, E., Ergo, K., Naert, L., Janssens, C., Talsma, D., Van Opstal, F., & Verguts, T. (2018). Signed reward prediction errors drive declarative learning. *PLoS One, 13*(1), e0189212. https://doi.org/10.1371/journal.pone.0189212

Dickinson, A., & Burke, J. (1996) Within compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology Section B, 49*(1b), 60-80, https://doi.org/10.1080/713932614

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *The Quarterly Journal of Experimental Psychology, 36*(1), 29-50. https://doi.org/10.1080/14640748408401502

Don, H. J., Beesley, T., & Livesey, E. J. (2019). Learned predictiveness models predict opposite attention biases in the inverse base-rate effect. *Journal of Experimental Psychology: Animal Learning and Cognition*, *45*(2), 143-162. https://doi.org/10.1037/xan0000196

Don, H. J., & Livesey, E. J. (2015). Resistance to instructed reversal of the learned predictiveness effect. *The Quarterly Journal of Experimental Psychology*, *68*(7), 1327-1347. https://doi.org/10.1080/17470218.2014.979212

Don, H. J. & Livesey, E. J. (2018). Is the blocking effect sensitive to causal model? It depends how you ask. *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp.1633–1638). Cognitive Science Society

Don, H. J., Worthy, D. A., & Livesey, E. J. (2021). Hearing hooves, thinking zebras: A review of the inverse base-rate effect. *Psychonomic Bulletin & Review, 28*, 1142–1163. https://doi.org/10.3758/s13423-020-01870-0

Enquist M., Ghirlanda S. (2005) Neural Networks and Animal Behavior. Princeton University Press.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*, 25-42. https://doi.org/10.1257/089533005775196732

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*, 127–138. https://doi.org/10.1038/nrn2787

Goldwater, M. B., Don, H. J., Krusche, M. J. F., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General, 147*(1), 1-35, https://doi.org/10.1037/xge0000387

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language, 94*, 149-165. https://doi.org/10.1016/j.jml.2016.11.001

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*(4), 334-384. https://doi.org/10.1016/j.cogpsych.2005.05.004

Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E. J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*(1), 123-138. https://doi.org/10.1080/23743603.2017.1326760

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E-J. (2021). A Primer on Bayesian Model-Averaged Meta-Analysis. *Advances in Methods and Practices in Psychological Science. 4*(3). https://doi.org/10.1177/25152459211031256

Hartanto, G., Livesey, E., Griffiths, O., Lachnit, H., & Thorwart, A. (2021). Outcome unpredictability affects outcome-specific motivation to learn. *Psychonomic Bulletin & Review, 28*(5), 1648-1656. https://doi.org/10.3758/s13423-021-01932-x

Heck, D. W., Gronau, Q. F., & Wagenmakers, E.-J. (2019). metaBMA: Bayesian model averaging for random and fixed effects meta-analysis. https://CRAN.R-project.org/package=metaBMA

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology, 62*, 135-163. https://doi.org/10.1146/annurev.psych.121208.131634

Hume, D. (1740/1978). *A treatise of human nature*. Oxford University Press.

Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction errors during decision-making strengthen memory encoding. *Nature Human Behaviour, 3*(7), 719-732. https://doi.org/10.1038/s41562-019-0597-3

JASP Team (2020). JASP (Version 0.12.2) [Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological monographs: General and applied, 79*(1), 1-17. https://doi.org/10.1037/h0093874

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*(1), 111-124. https://doi.org/10.1037/xan0000185

Kamin, L. J. (1968). Attention-like processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9-32). University of Miami Press.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception,* 36:ECVP Abstract Supplement.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of mathematical psychology, 45*(6), 812-863. https://doi.org/10.1006/jmps.2000.1354

Kruschke, J. K. (2003). Attention in Learning. *Current Directions in Psychological Science, 12*(5), 171-175. https://doi.org/10.1111/1467-8721.01254

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636-645. https://doi.org/10.3758/BF03213001

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 830-845. https://doi.org/10.1037/0278-7393.31.5.830

Larkin, M. J. W., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1331–1352. https://doi.org/10.1037/0278-7393.24.6.1331

Lee Cheong Lem, V. A., Moul, C., Harris, J. A., & Livesey, E. J. (2018). Associations or repetitions? Testing the basis of the Perruchet effect in voluntary response speed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1971-1985. https://doi.org/10.1037/xlm0000556

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*(3b), 193-243. https://doi.org/10.1080/02724990344000141

Le Pelley, M. E., Beesley, T., & Suret, M. B. (2007). Blocking of human causal learning involves learned changes in stimulus processing. *Quarterly journal of experimental psychology, 60*(11), 1468-1476. https://doi.org/10.1080/17470210701515645

Le Pelley, M. E., Griffiths, O., & Beesley, T. (2017). Associative accounts of causal cognition. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 13-28). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199399550.001.0001

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology Section B, 56*(1b), 68-79. https://doi.org/10.1080/02724990244000179

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin, 142*(10), 1111–1140. https://doi.org/10.1037/bul0000064

Le Pelley, M. E., Pearson, D., Griffiths, O., & Beesley, T. (2015). When goals conflict with values: counterproductive attentional and oculomotor capture by reward-related stimuli. *Journal of Experimental Psychology: General, 144*(1), 158. https://doi.org/10.1037/xge0000037

Le Pelley, M. E., Vadillo, M., & Luque, D. (2013). Learned predictiveness influences rapid attentional capture: evidence from the dot probe task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1888. https://doi.org/10.1037/a0033700

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron, 46*(5), 703-713. https://doi.org/10.1016/j.neuron.2005.05.002

Livesey, E. J. (2023). Computational models of animal and human associative learning. In R. Sun (Ed.) *The Cambridge Handbook on Computational Cognitive Sciences* (2nd ed., pp. 703-738). Cambridge University Press. https://doi.org/10.1017/9781108755610.025

Livesey, E. J., Greenaway, J. K., Schubert, S., & Thorwart, A. (2019). Testing the deductive inferential account of blocking in causal learning. *Memory & Cognition.*, *47*(6), 1120–1132. https://doi.org/10.3758/s13421-019-00920-w

Livesey, E. J., Lee, J., & Shone, L. (2013). The relationship between blocking and inference in causal learning. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2920–2925).

Lober, K. & Shanks, D. R. (2000). Is Causal Induction Based on Causal Power? Critique of Cheng (1997). *Psychological Review, 107*(1), 195-212. https://doi.org/10.1037/0033-295X.107.1.195

López, F. J., Cobos, P. L., & Caño, A. (2005). Associative and causal reasoning accounts of causal induction: Symmetries and asymmetries in predictive and diagnostic inferences. Memory & Cognition, 33(8), 1388-1398. https://doi.org/10.3758/BF03193371

Lovibond, P. F. (2003). Causal Beliefs and Conditioned Responses: Retrospective Revaluation Induced by Experience and by Instruction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29* (1), 97-106. https://doi.org/10.1037/0278-7393.29.1.97

Lovibond, P.F., 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 & Cognition*, *31*(1), 133-142. https://doi.org/10.3758/BF03196088

Luque, D., Vadillo, M. A., Gutiérrez-Cobo, M. J., & Le Pelley, M. E. (2018). The blocking effect in associative learning involves learned biases in rapid attentional capture. *Quarterly Journal of Experimental Psychology*, *71*(2), 522-544. https://doi.org/10.1080/17470218.2016.1262435

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological review, 82*(4), 276. https://doi.org/10.1037/h0076778

Matute, H., Arcediano, F., & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 182–196. https://doi.org/10.1037/0278-7393.22.1.182

Matute, H., Blanco, F., & Díaz-Lago, M. (2019). Learning mechanisms underlying accurate and biased contingency judgments. *Journal of Experimental Psychology: Animal Learning and Cognition, 45*(4), 373–389. https://doi.org/10.1037/xan0000222

McAndrew, A., Jones, F. W., McLaren, R. P., & McLaren, I. P. L. (2012). Dissociating expectancy of shock and changes in skin conductance: An investigation of the Perruchet effect using an electrodermal paradigm. *Journal of Experimental Psychology: Animal Behavior Processes, 38*(2), 203 –208. https://doi.org/10.1037/a0026718

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed Memory and the Representation of General and Specific Information. *Journal of Experimental Psychology: General, 114*(2), 159-188. https://doi.org/10.1037/0096-3445.114.2.159

McLaren, I. P., Forrest, C. L. D., McLaren, R. P., Jones, F. W., Aitken, M. R. F., & Mackintosh, N. J. (2014). Associations and propositions: The case for a dual-process account of learning in humans. *Neurobiology of learning and memory, 108*, 185-195. https://doi.org/10.1016/j.nlm.2013.09.014

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior, 30*, 177-200. https://doi.org/10.3758/BF03192828

McNally, R. J. (1981). Phobias and preparedness: instructional reversal of electrodermal conditioning to fear-relevant stimuli. *Psychological reports, 48*(1), 175-180. https://doi.org/10.2466/pr0.1981.48.1.175

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioural and Brain Sciences, 32*, 183–198. https://doi.org/10.1017/S0140525X09000855

Mitchell, C. J., Griffiths, O., Seetoo, J., & Lovibond, P. F. (2012). Attentional mechanisms in learned predictiveness. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(2), 191-202. https://doi.org/10.1037/a0027385

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 Section B, 55*(4b), 311-329. https://doi.org/10.1080/02724990244000025

Mitchell, C. J., Lovibond, P. F., & Gan, C. Y. (2005). A dissociation between causal judgment and outcome recall. *Psychonomic bulletin & review*, *12*, 950-954. https://doi.org/10.3758/BF03196791

Mitchell, C. J., Lovibond, P. F., Minard, E., & Lavis, Y. (2006). Forward blocking in human learning sometimes reflects the failure to encode a cue–outcome relationship. *The Quarterly Journal of Experimental Psychology*, *59*, 830-844. https://doi.org/10.1080/17470210500242847

Morey, R. (2015). Multiple Comparisons with BayesFactor, Part 1. *BayesFactor: An R package for Bayesian data analysis. http://bayesfactor.blogspot.com/2015/01/multiple-comparisons-with-bayesfactor-1.html.*

Morey, R. D. & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Morís, J., Cobos, P. L., Luque, D., and López, F. J. (2014). Associative repetition priming as a measure of human contingency learning: evidence of forward and backward blocking. *Journal of Experimental Psychology General, 143*(1), 77–93. https://doi.org/10.1037/a0030919

Nasser, H. M., Calu, D. J., Schoenbaum, G., and Sharpe, M. J. (2017). The Dopamine Prediction Error: Contributions to Associative Models of Reward Learning. *Frontiers in Psychology 8*, 244. https://doi.org/10.3389/fpsyg.2017.00244

Pavlov, I. P. (1927). *Conditioned Reflexes.* Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, *87*(6), 532. https://doi.org/10.1037/0033-295X.87.6.532

Pearce, J. M., & Mackintosh, N. J. (2010). Two theories of attention: A review and a possible integration. In C. J. Mitchell & M. E. LePelley (Eds.), *Attention and associative learning: From brain to behaviour* (pp. 11–40). Oxford: Oxford University Press.

Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. *The Pavlovian journal of biological science*, *20*(4), 163-170. https://doi.org/10.1007/BF03003653

Press, C., Kok, P., & Yon, D. (2020). The perceptual prediction paradox. *Trends in Cognitive Sciences, 24*(1), 13-24. https://doi.org/10.1016/j.tics.2019.11.003

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of comparative and physiological psychology, 66*(1), 1-5. https://doi.org/10.1037/h0025984

Rescorla, R. A. (1972). Informational variables in Pavlovian conditioning. In G.H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 6, pp. 1-46). Academic Press.

Rescorla, R. A., & 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* (pp. 64–99). Appleton-Century-Crofts.

Rouder, J. N., Morey, R. D., Speckman, P. L., Province, J. M., (2012) Default Bayes Factors for ANOVA Designs. *Journal of Mathematical Psychology, 56*, 356-374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Engelhardt C. R., McCabe S., & Morey R. D. (2016). Model comparison in ANOVA. *Psychonomic Bulletin and Review, 23*, 1779-1786. https://doi.org/10.3758/s13423-016-1026-5

Rouder, J. N., Morey, R. D., Verhagen, A. J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22*, 304-321. https://doi.org/10.1037/met0000057

Sahakyan, L., Delaney, P. F., Foster, N. L., & Abushanab, B. (2013). List-method directed forgetting in cognitive and clinical research: A theoretical and methodological review. *Psychology of learning and motivation, 59*, 131-189. https://doi.org/10.1016/B978-0-12-407187-2.00004-6

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications* (pp. 71-78).

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1-21. https://doi.org/10.1080/14640748508402082

Shanks, D. R. (1991). On similarities between causal judgments in experienced and described situations. *Psychological Science, 2*(5), 341-350. https://doi.org/10.1111/j.1467-9280.1991.tb00163.x

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *The Quarterly Journal of Experimental Psychology*, *60*(3), 291-309. https://doi.org/10.1080/17470210601000581

Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition, 19,* 353-360. https://doi.org/10.3758/BF03197139

Shanks, D. R., & Lopez, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory & Cognition, 24*(4), 511-522. https://doi.org/10.3758/BF03200939

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., Niv, Y., & Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience, 20*(5), 735-742. https://doi.org/10.1038/nn.4538

Shone, L. T., Harris, I. M., & Livesey, E. J. (2015). Automaticity and cognitive control in the learned predictiveness effect. *Journal of Experimental Psychology: Animal Learning and Cognition, 41*, 18–31. https://doi.org/10.1037/xan0000047

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin, 119*(1), 3-22. https://doi.org/10.1037/0033-2909.119.1.3

Spicer, S. G., Mitchell, C. J., Wills, A. J., Blake, K. L., & Jones, P. M. (2022). Theory protection: Do humans protect existing associative links*? Journal of Experimental Psychology: Animal Learning and Cognition, 48*(1), 1–16. https://doi.org/10.1037/xan0000314

Spicer, S. G., Mitchell, C. J., Wills, A. J., & Jones, P. M. (2020). Theory protection in associative learning: Humans maintain certain beliefs in a manner that violates prediction error. *Journal of Experimental Psychology: Animal Learning and Cognition, 46*(2), 151–161. https://doi.org/10.1037/xan0000225

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of cognitive neuroscience, 4*(3), 232-243. https://doi.org/10.1162/jocn.1992.4.3.232

Stalnaker, T. A., Howard, J. D., Takahashi, Y. K., Gershman, S. J., Kahnt, T., & Schoenbaum, G. (2019) Dopamine neuron ensembles signal the content of sensory prediction errors. *eLife, 8*, e49315. https://doi.org/10.7554/eLife.49315

Stanek, J. K., Dickerson, K. C., Chiew, K. S., Clement, N. J., & Adcock, R. A. (2019). Expected reward value and reward uncertainty have temporally dissociable effects on memory formation. *Journal of Cognitive Neuroscience, 31*(10), 1443-1454. https://doi.org/10.1162/jocn_a_01411

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. Journal of Experimental Psychology: General, 127(2), 161–188. https://doi.org/10.1037/0096-3445.127.2.161

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience, 16*(7), 966-973. https://doi.org/10.1038/nn.3413

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*(2), 135-170. https://doi.org/10.1037/0033-295X.88.2.135

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron, 95*(6), 1395-1405. https://doi.org/10.1016/j.neuron.2017.08.025

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 35–42). MIT Press.

Thorwart, A., & Livesey, E. J. (2016). Three ways that non-associative knowledge may affect associative learning processes. *Frontiers in psychology*, *7*, 2024. https://doi.org/10.3389/fpsyg.2016.02024

Thorwart, A., Livesey, E. J., & Harris, J. A. (2012). Normalization between stimulus elements in a model of Pavlovian conditioning: Showjumping on an elemental horse. *Learning & behavior, 40,* 334-346. https://doi.org/10.3758/s13420-012-0073-7

Tran, D. M. D., Harris, J. A., Harris, I. M., & Livesey, E. J. (2020). Motor conflict: Revealing involuntary conditioned motor preparation using transcranial magnetic stimulation. *Cerebral Cortex, 30*(4), 2478-2488. https://doi.org/10.1093/cercor/bhz253

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124-1131.

Vandorpe, S., & De Houwer, J. (2005). A comparison of forward blocking and reduced overshadowing in human causal learning. *Psychonomic Bulletin & Review*, *12*, 945-949. https://doi.org/10.3758/BF03196790

Van Hamme, L. J., Kao, S. F., & Wasserman, E. A. (1993). Judging interevent relations: From cause to effect and from effect to cause. *Memory & Cognition, 21*(6), 802-808. https://doi.org/10.3758/BF03202747

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*(2), 127–151. https://doi.org/10.1006/lmot.1994.1008

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412*, 43–48. https://doi.org/10.1038/35083500

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), Information processing in animals: Memory mechanisms (pp. 5–47). Erlbaum.

Wagenmakers, E.J., Love, J., Marsman, M. et al. (2018) Bayesian inference for psychology. Part II: example applications with JASP. *Psychonomic Bulletin and Review, 25,* 58-76. https://doi.org/10.3758/s13423-017-1323-7

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53-76. https://doi.org/10.1037/0278-7393.26.1.53

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review*, *8*(3), 600-608. https://doi.org/10.3758/BF03196196

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222-236. https://doi.org/10.1037/0096-3445.121.2.222

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology/Revue canadienne de psychologie, 19*(3), 231-241. https://doi.org/10.1037/h0082908

Wasserman, E. A. (1990). Attribution of causality to common and distinctive elements of compound stimuli. *Psychological Science, 1*(5), 298-302. https://doi.org/10.1111/j.1467-9280.1990.tb00221.x

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92-107. https://doi.org/10.1006/jmps.1999.1278

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and motivation, 14*(4), 406-432. https://doi.org/10.1016/0023-9690(83)90025-5

Weidemann, G., McAndrew, A., Livesey, E. J., & McLaren, I. P. (2016). Evidence for multiple processes contributing to the Perruchet effect: Response priming and associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition, 42*(4), 366. https://doi.org/10.1037/xan0000117

Wheeler, D. S., & Miller, R. R. (2007). Contrasting reduced overshadowing and blocking. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(3), 349-359. https://doi.org/10.1037/0097-7403.33.3.349

Widrow, B., & Hoff, M. E. (1960, August). Adaptive switching circuits. In *IRE WESCON convention record* (Vol. 4, No. 1, pp. 96-104).

Wills, A. J., Lavric, A., Croft, G. S., & Hodgson, T. L. (2007). Predictive learning, prediction errors, and attention: Evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience*, *19*(5), 843-854. https://doi.org/10.1162/jocn.2007.19.5.843

# Appendix A: Experimental task instructions

## Chapter 2

**Introductory instructions**

"This experiment is designed to help us understand how people make judgements about cause and effect. Please note that this is not a test of your ability levels and that all information will be collected anonymously. You will now be presented with a set of instructions for completing the experiment. If at any time anything is unclear, you are free to ask questions.

In this experiment you are asked to imagine you are an allergist (someone who tries to discover the cause of allergic reactions in people). A new patient, Miss Y/Mr X,[1] has come to you because she/he is suffering from several allergic reactions after eating certain foods. In an attempt to discover which foods cause the different types of allergic reactions, you arrange for Miss Y/Mr X to eat a number of different meals, containing one or two foods, and observe the type of allergic reaction she/he suffers.

In a moment you will see each of the results from these tests one by one. Note that Miss Y/Mr X suffers allergic reactions very frequently after eating foods, but she/he does not always suffer the same reaction. Your task is to learn which foods are causing which reactions. Pay close attention to what foods were eaten, and which allergic reaction followed.

Please note that only the presented information can help you. Your own personal knowledge or experience with food allergies will NOT help you in this task. Specifically, you need to decide which foods are causing an allergic reaction in THIS particular patient. Consequently, try to use only the knowledge you have learned from the experiment to make your decisions."

**Pretraining instructions**

[*EXPERIMENT 2.2 ONLY: "At the beginning of each trial, a small cross will appear in the centre of the screen. It is important that you fix your eyes on this cross while it is on the screen."*[2]*]*

**"**You will now be shown the contents of meals eaten by Miss Y/Mr X, and will be asked to predict what outcome will occur as a result of eating that meal. You will have a choice of two different kinds of allergic reactions or no allergic reaction.

Use the mouse to indicate which allergic reaction you think is more likely to occur after eating these foods. You will then be told whether your prediction was correct or incorrect.

---

[1] All experiments in chapter two included a manipulation of the number of patients. To achieve this, the initial patient name was manipulated such that those in the same patient group saw only Miss Y throughout the experiment, whereas those in the new patient group saw Mr X in pretraining and Miss Y in the final two phases.

[2] These instructions appeared only in Experiment 2.2 as it included eye-tracking and therefore required a reorientation of attention towards the centre of the screen before every trial.

At first you will merely have to guess which allergic reaction is going to occur. But using the feedback provided, you should find that your predictions improve over time. If you have any questions, then please ask the experimenter now."

**Training instructions**

**Same Patient Condition.** (Pretraining patient Miss Y)

**"**Please take a rest for a few moments. When you restart, you will once again be presented with a set of specific foods that Miss Y has eaten. Some of the foods given to Miss Y are the same as before, some are not. Again, your task is to predict which allergic reaction will occur. You will be given feedback about the accuracy of your response after each prediction. Please use this feedback to guide future responses."

**New Patient Condition.** (Pretraining patient Mr X)

**"**Now, another patient has come to visit you. Miss Y has come to visit you because she is also suffering allergic reactions after eating certain foods. You will now be presented with a set of foods that Miss Y has eaten. Some of the foods given to Miss Y are the same as those given to Mr X, some are not. As in the previous phase, your task is to predict which allergic reaction will occur. You will be given feedback about the accuracy of your response after each prediction. Please use this feedback to guide future responses."

**Test instructions**

**Experiments 2.1, 2.2, and 2.3b.**

"The next part of the experiment requires you to use the knowledge you have gained so far. You will be shown a number of foods eaten by Miss Y, and will be asked to make two judgements about each food. First, indicate which reaction occurred after Miss Y ate this food. This is a bit like a memory test to see how well you can recall which foods and reactions were presented together.

Three rating scales will appear, one for each type of allergic reaction experienced by Miss Y. Use the mouse to click on the point on each scale that reflects your confidence that this food was followed by Fever, Nausea and No reaction. Once you have made all three ratings, press the space bar to continue to the second part.

Not every food caused an allergic reaction in Miss Y. However, each food was CONSISTENTLY followed by just ONE outcome in Miss Y. Your task is to recall which outcome followed each food.

Second, indicate to what degree the food caused an allergic reaction of any kind in Miss Y[3]. A rating scale will appear ranging from "Definitely DOES NOT cause a reaction" to "Definitely DOES cause a reaction"

Use the mouse to click on the point on the scale that reflects your confidence that this food will cause an allergic reaction. Once you have made both your ratings, press the space bar to continue to the next trial. If you have any questions, then please ask the experimenter now."

***On each test trial**:*

*Memory Scales*: "Which allergic reaction followed after Miss Y ate this food?"

*Causal Scale*: "How likely is it that this food CAUSES an allergic reaction in Miss Y?"

**Experiment 2.3a.**

***If memory test first:*** "The next part of the experiment requires you to use the knowledge you have gained so far. You will be shown a number of foods eaten by Miss Y and will be asked to make a judgement about each food.

Please indicate which allergic reaction followed after Miss Y ate each food. This is a bit like a memory test to see how well you can recall which foods and reactions were presented together. Three rating scales will appear, one for each type of allergic reaction experienced by Miss Y. Use the mouse to click on the point on each scale that reflects your confidence that this food was followed by Fever, Nausea and No reaction. Once you have made all three ratings, press the space bar to continue to the next food.

Not every food caused an allergic reaction in Miss Y. However, each food was CONSISTENTLY followed by just ONE outcome in Miss Y. Your task is to recall which outcome followed each food.

[Memory test block completed]

Well done, you have completed the memory test. We will now ask you to make a different set of judgements about the foods that Miss Y has eaten.

Again, you will be shown a number of foods eaten by Miss Y. However, this time your task is to indicate to what degree the food caused an allergic reaction of ANY kind in Miss Y. A rating scale will appear ranging from ''Definitely DOES NOT cause a reaction'' to ''Definitely DOES cause a reaction''. Use the mouse to click on the point on the scale that reflects your confidence that this food will cause an allergic reaction. Once you have made all three ratings, press the space bar to continue to the next food."

[3] Note that in Experiment 2.3b the order of the test judgements was counterbalanced, such that half of the participants were asked to make the causal judgements before the memory judgements. The test instructions they received were identical aside from the order in which the judgements were described.

***If causal ratings first:*** **"**The next part of the experiment requires you to use the knowledge you have gained so far. You will be shown a number of foods eaten by Miss Y and will be asked to make a judgement about each food.

Your task is to indicate to what degree the food caused an allergic reaction of ANY kind in Miss Y. A rating scale will appear ranging from ''Definitely DOES NOT cause a reaction'' to ''Definitely DOES cause a reaction''. Use the mouse to click on the point on the scale that reflects your confidence that this food will cause an allergic reaction. Once you have made all three ratings, press the space bar to continue to the next food.

[Causal test block completed]

Well done, you have completed the causal ratings. We will now ask you to make a different set of judgements about the foods that Miss Y has eaten.

Again, you will be shown a number of foods eaten by Miss Y. However, this time your task is to indicate which allergic reaction followed after Miss Y ate each food. This is a bit like a memory test to see how well you can recall which foods and reactions were presented together.

Three rating scales will appear, one for each type of allergic reaction experienced by Miss Y. Use the mouse to click on the point on each scale that reflects your confidence that this food was followed by Fever, Nausea and No reaction. Once you have made all three ratings, press the space bar to continue to the next food.

Not every food caused an allergic reaction in Miss Y. However, each food was CONSISTENTLY followed by just ONE outcome in Miss Y. Your task is to recall which outcome followed each food."

**Survey of allergy assumptions (Experiment 2.3b only)**

"Finally, we would like to ask you to read a short scenario about food allergies and rate the extent to which you agree with each of three statements made about the circumstances outlined in the scenario.

> **SCENARIO**: Imagine that you are an allergist, and you discover that one of your patients (Patient 1) is allergic to grapes such that they suffer a headache whenever they eat grapes.
>
> Then, another patient (Patient 2) comes to see you because they are also suffering from allergic reactions after eating certain foods.
>
> How would the information that Patient 1 is allergic to grapes influence your assessment of the likelihood that Patient 2 is allergic to grapes?

Please indicate the extent to which you agree with the statement below on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree).

Use the number keys at the top of the keyboard to make a response."

[The above scenario was presented on each screen and participants were asked to respond to each of the three statements below one at a time in a randomised order.]

**STATEMENT 1:** 'If Patient 1 suffers a headache whenever they eat grapes, Patient 2 is more likely to suffer a headache whenever they eat grapes.'

**STATEMENT 2:** 'If Patient 1 suffers a headache whenever they eat grapes, Patient 2 is more likely to likely to be allergic to grapes even if the specific symptom they suffer is different to that of Patient 1.'

**STATEMENT 3:** 'If Patient 1 is allergic to grapes that would not increase the likelihood that Patient 2 is allergic to grapes.'

**Chapter 3**

Introductory, pretraining, and test phase[4] instructions were identical to those given to the new patient group of Experiment 2.1, instructions for the training phase were modified as below.

**Training instructions**

**New Patient Neutral Condition. "**Now, another patient has come to visit you. Miss Y has come to visit you because she is also suffering allergic reactions after eating certain foods. You will now be presented with a set of foods that Miss Y has eaten. Some of the foods given to Miss Y are the same as those given to Mr X, some are not.

As in the previous phase, your task is to predict which allergic reaction will occur. You will be given feedback about the accuracy of your response after each prediction. Please use this feedback to guide future responses."

**New Patient Known Condition. "**Now, another patient has come to visit you. Miss Y has come to visit you because she is also suffering allergic reactions after eating certain foods. You will now be presented with a set of foods that Miss Y has eaten. Some of the foods given to Miss Y are the same as those given to Mr X, some are not.

[4] Note that in Experiment 3.1 the 'no allergic reaction' outcome was not included in the design and the test procedure was changed. Accordingly, the test instructions from Experiment 2.1 describing how to make memory ratings were modified in Experiment 3.1 with the following text "A rating scale will appear with options corresponding to the two allergic reactions suffered by Miss Y. Use the mouse to click on the point on the scale that reflects your confidence that this food was followed by 'Nausea' or 'Fever.'"

As in the previous phase, your task is to predict which allergic reaction will occur. You will be given feedback about the accuracy of your response after each prediction. Please use this feedback to guide future responses.

The following information is ALREADY KNOWN about Miss Y's allergies: Miss Y suffers from NAUSEA after eating DAIRY such as YOGHURT, CHEESE, MILK, and ICECREAM

AND

Miss Y suffers from FEVER after eating MEAT such as BACON, BEEF, CHICKEN, and FISH"[5]

## Chapter 4

### Experiment 4.1

Introductory, pretraining, and test instructions were identical to those used in chapter 2, the pretraining patient was Mr X in all three conditions. The training instructions are given below.

**New Patient Condition. "**Now, another patient has come to visit you. Miss Y has come to visit you because she is also suffering allergic reactions after eating certain foods. You will now be presented with a set of foods that Miss Y has eaten. Some of the foods given to Miss Y are the same as those given to Mr X, some are not.

As in the previous phase, your task is to predict which allergic reaction will occur. You will be given feedback about the accuracy of your response after each prediction. Please use this feedback to guide future responses."

**Retroactive Practice Condition. "**You have completed the first part of the experiment. However, the patient you have just seen was a practice example to give you some experience with the task. You will NOT be tested on what you have learned about Mr X's allergies.

Now the real task will begin. You will be presented with a set of specific foods that a different patient, Miss Y, has eaten. Some of the foods given to Miss Y are the same as before, some are not.

As in the practice phase, you need to predict which allergic reaction will occur after Miss Y has eaten each meal. You will be given feedback about the accuracy of your response after each prediction. Please use this feedback to guide future responses. Try to remember this information throughout the next phase. Your task is to determine which of these foods cause which of the specific allergies that Miss Y suffers from. You will be tested on this afterwards."

[5] This is an example of the instruction given, the mapping of symptom to foods and food category was randomised by participant.

**Proactive Practice Condition.** (Before pretraining began) “PRACTICE: The first patient you will see is just for practice to give you some experience with the task. You will NOT be tested on what you learn about Mr X’s allergies.”

[Pretraining completed]

“You have completed the practice trials! Now the real task will begin. You will be presented with a set of specific foods that a different patient, Miss Y, has eaten. Some of the foods given to Miss Y are the same as before, some are not.

As in the practice phase, you need to predict which allergic reaction will occur after Miss Y has eaten each meal. You will be given feedback about the accuracy of your response after each prediction. Please use this feedback to guide future responses. Try to remember this information throughout the next phase. Your task is to determine which of these foods cause which of the specific allergies that Miss Y suffers from. You will be tested on this afterwards.”

**Experiment 4.2**

Introductory instructions were modified to accommodate the change from in person to online testing. The full script along with the instruction check test are given below.

“Thanks for accepting our [SONA/Prolific Academic] advertisement. This is a short psychological study investigating how people learn about cause and effect and make predictions. It involves the following steps:

1. We ask for demographic information (not connected to your [student/personal ID])
2. The study then explains how to do the task in detail. You will need to pass a short test to check that you understand how the study works.
3. Next comes the experiment itself.
4. At the end, [we will give you the completion code you need to get credit for SONA participation/you will be redirected to Prolific Academic].

The total time taken should be about 30 minutes.

Please **do not** use the "back" button on your browser or close the window until you reach the end and [receive your completion code/are redirected to the Prolific Academic website]. This is very likely to break the experiment.

It is also important that you complete the task in **one** sitting on a **computer**. If you attempt to complete the experiment on another device (e.g. tablet, phone), then you may not be able to complete the experiment.

Please note that you will not be able to scroll up and down on some of the screens presented in this experiment. If you cannot see all the relevant material on your screen, then try maximising your browser window. You can also select *zoom out* in the View menu of your browser.

If something does go wrong, please contact us! When you are ready to begin, click on the "Start" button below."

[Consent and demographic info collected]

"This experiment is designed to help us understand how people make judgements about cause and effect. Please note that all information will be collected anonymously.

You will now be presented with a set of instructions for completing the experiment. Please read them CAREFULLY.

In this experiment you are asked to imagine that you are an allergist (someone who tries to discover the cause of allergic reactions in people). A new patient, (Mr X/Miss Y), has come to you because she is suffering from several allergic reactions after eating certain foods. In an attempt to discover which foods, cause the different types of allergic reactions, you arrange for Miss Y to eat a number of different meals, each containing one or two foods, so that you can observe the type of allergic reaction she suffers.

In a moment you will see each of the results from these tests one by one. Note that Miss Y suffers allergic reactions very frequently after eating foods, but she does not always suffer the same reaction. Your task is to learn which foods are causing which reactions. Pay close attention to what foods were eaten, and which allergic reaction followed.

Please note that only the presented information can help you. Your own personal knowledge or experience with food allergies will NOT help you in this task. Specifically, you need to decide which foods are causing an allergic reaction in THIS particular patient. Consequently, try to use only the knowledge you have learned from the experiment to make your decisions.

You will now be shown the contents of meals eaten by Miss Y, and will be asked to predict what kind of allergic reaction will result from eating each meal. Use the mouse to indicate which allergic reaction you think is more likely to occur after eating these foods. You will then be told whether your prediction was correct or incorrect. At first you will merely have to guess which allergic reaction is going to occur. But using the feedback provided, you should find that your predictions improve over time.

Please DO NOT write anything down during the experiment. It will affect the results and will slow you down.

Press Next to begin the experiment."

**Instruction check survey.** Participants were required to answer all four items correctly to proceed to the experiment. They were given unlimited attempts. If they answered any question incorrectly, they were asked to re-read the instructions before making another attempt. Any participant who failed the survey more than two times was excluded from the analyses.

**Check your knowledge before you begin!**

**Question 1:** The aim of this task is to determine which foods cause your patient to suffer from particular allergic reactions:*

- TRUE
- FALSE

**Question 2:** On each trial, your task will be to make a prediction about which allergic reaction your patient will suffer given that they have eaten certain foods. Later in the experiment, you will be asked questions about which allergic reactions your patient suffered and which foods cause those allergic reactions.*

- TRUE
- FALSE

**Question 3:** You will have to do this through a process of trial and error, but you will receive feedback to help you learn. You will see the same combinations of foods presented multiple times.*

- TRUE
- FALSE

**Question 4:** Should I write down things as I go through this task? *

- Yes, sure it is fine to write things down.
- No, please DO NOT write down anything. It will affect the results and will slow you down.

**Training instructions.**

**Same patient + practice "**You have completed the first part of the experiment. However, the information you have just seen was a practice example to give you some experience with the task. You will NOT be tested on what you have learned about Miss Y's allergies. Please disregard what you have learned so far and remember the information you will be presented with throughout the next phase (it is this new information that you will be tested on afterwards).

Now the real task will begin. Again, you will be presented with a set of specific foods that your patient, Miss Y has eaten. Some of the foods given to Miss Y are the same as before, some are not. As in the practice phase, you need to predict which allergic reaction will occur after Miss Y has eaten each meal."

**New patient + practice** "You have completed the first part of the experiment. However, the patient you have just seen was a practice example to give you some experience with the task. You will NOT be tested on what you have learned about Mr X's allergies. Please disregard what you have learned so far and remember the information you will be presented with throughout the next phase (it is this new information will be tested on afterwards).

Now the real task will begin. You will be presented with a set of specific foods that a different patient, Miss Y, has eaten. Some of the foods given to Miss Y are the same as before, some are not. As in the practice phase, you need to predict which allergic reaction will occur after Miss Y has eaten each meal."

**Experiment 4.3**

Introductory and test instructions were the same as those used in chapter 2. The pretraining and training instructions were modified to reflect the change in procedure from active to passive learning. The modified instructions are given below.

**Pretraining instructions.** You will now be shown the contents of meals eaten by Miss Y/Mr X, and will observe the type of allergic reaction, if any, that Miss Y /Mr X suffers from after eating each meal. Use the mouse to click CONTINUE to see what happens when your patient eats each meal. If you have any questions, then please ask the experimenter now.

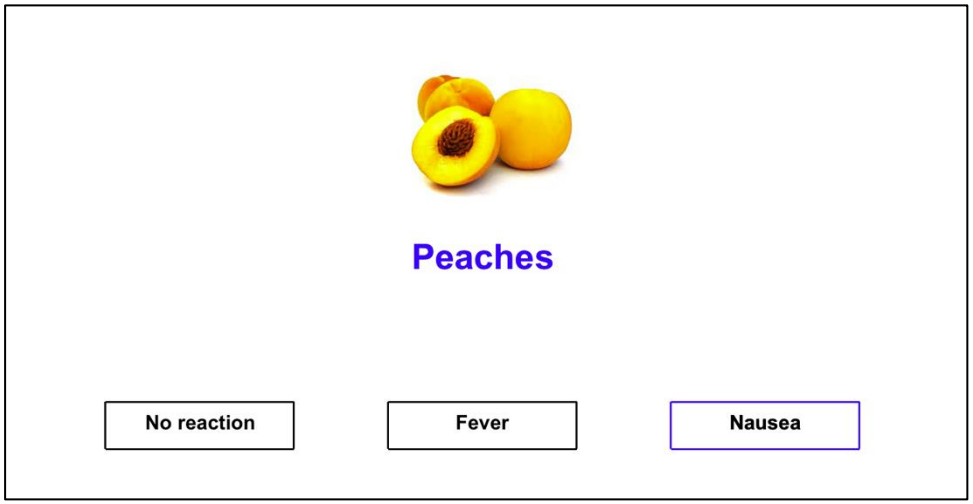
**Training instructions.**

***Same Patient Condition*** (Pretraining patient Miss Y) Please take a rest for a few moments. When you restart, you will once again be presented with a set of specific foods that Miss Y has eaten. Some of the foods given to Miss Y are the same as before, some are not. Again, your task is to learn which allergic reaction will occur. You will be tested on this afterwards.

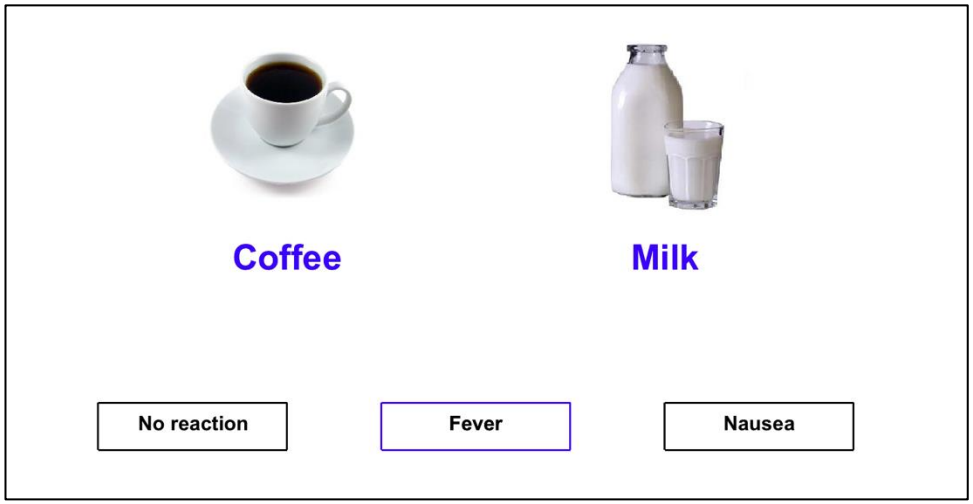
***New Patient Condition*** (Pretraining patient Mr X) Now, another patient has come to visit you. Miss Y has come to visit you because she is also suffering allergic reactions after eating certain foods. You will now be presented with a set of foods that Miss Y has eaten. Some of the foods given to Miss Y are the same as those given to Mr X, some are not. As in the previous phase, your task is to learn which allergic reaction will occur. You will be tested on this afterwards.

## Appendix B: Experimental task screenshots

Example of a typical training trial with single cue



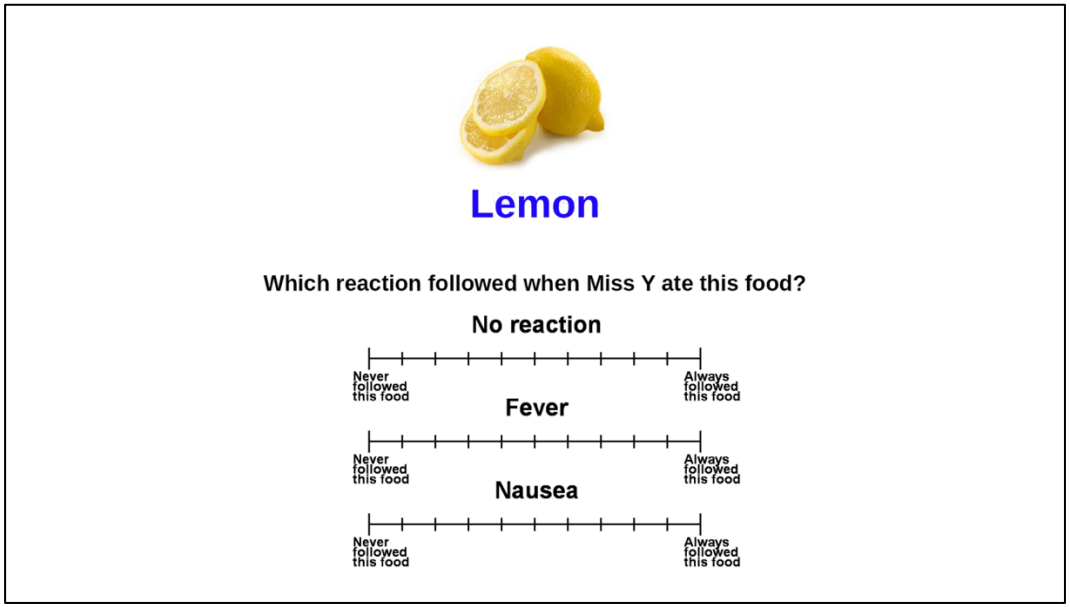

Example of a typical training trial with two cues



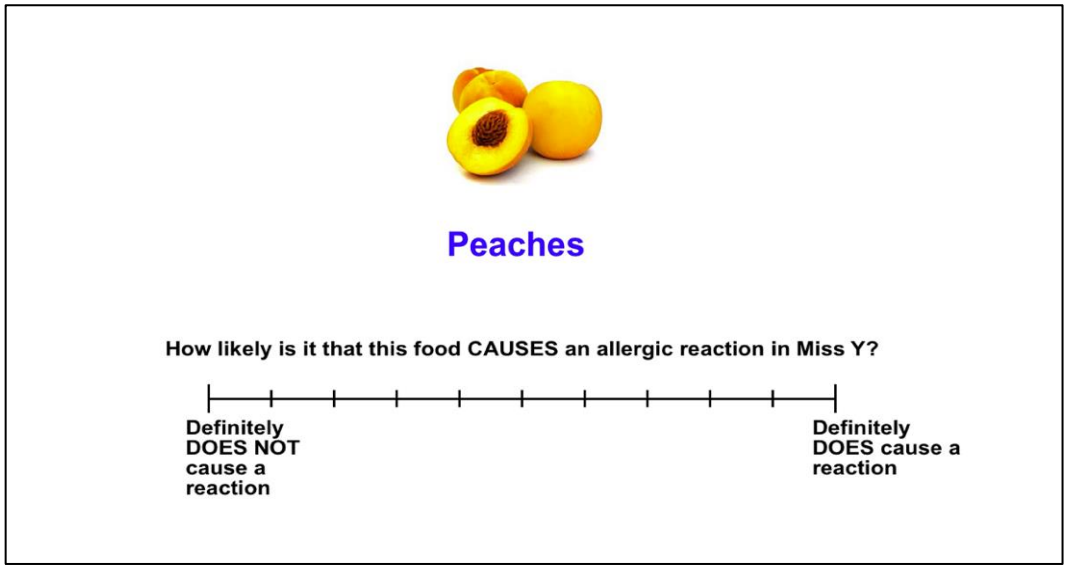

Example of corrective feedback for a correct response


Example of corrective feedback for an incorrect response

Example of a typical memory test trial

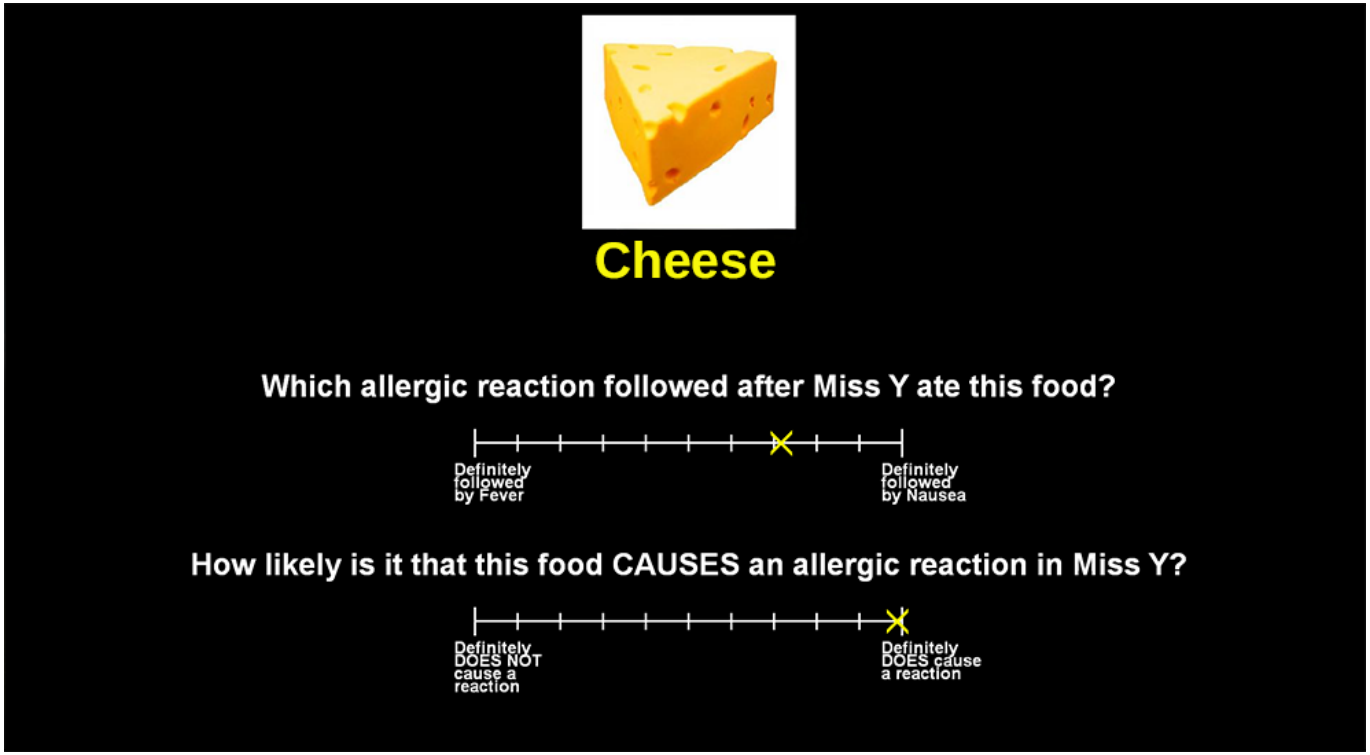



Example of a typical causal ratings test trial



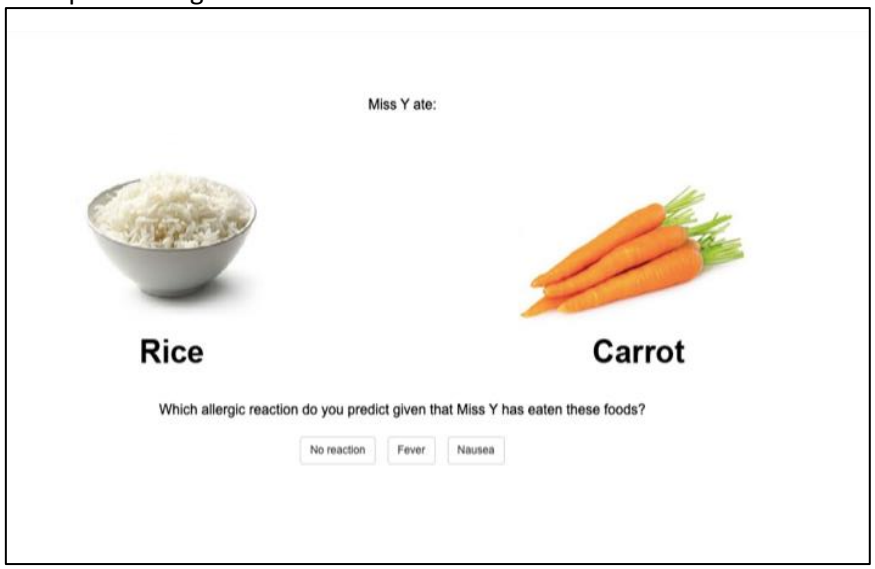
**Experiment 3.1**

Example of a test trial

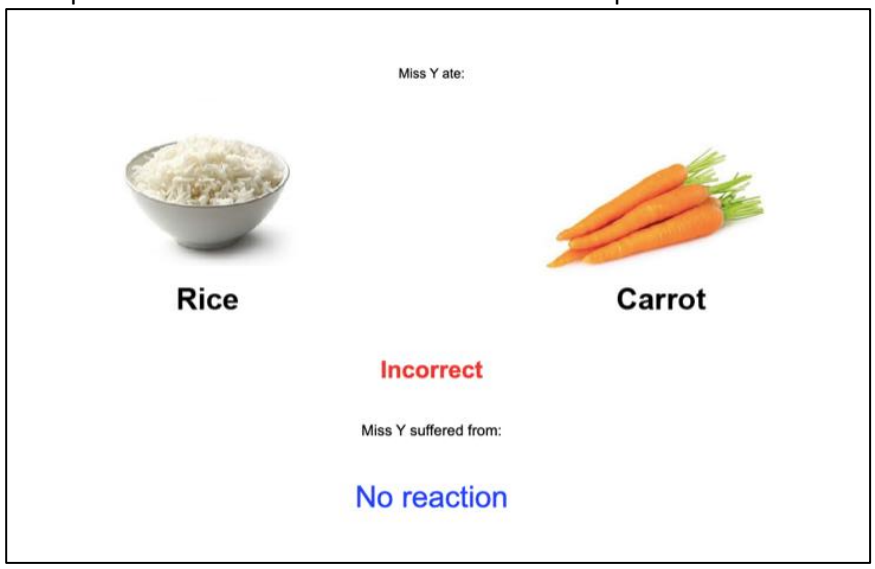


**Experiment 4.2**

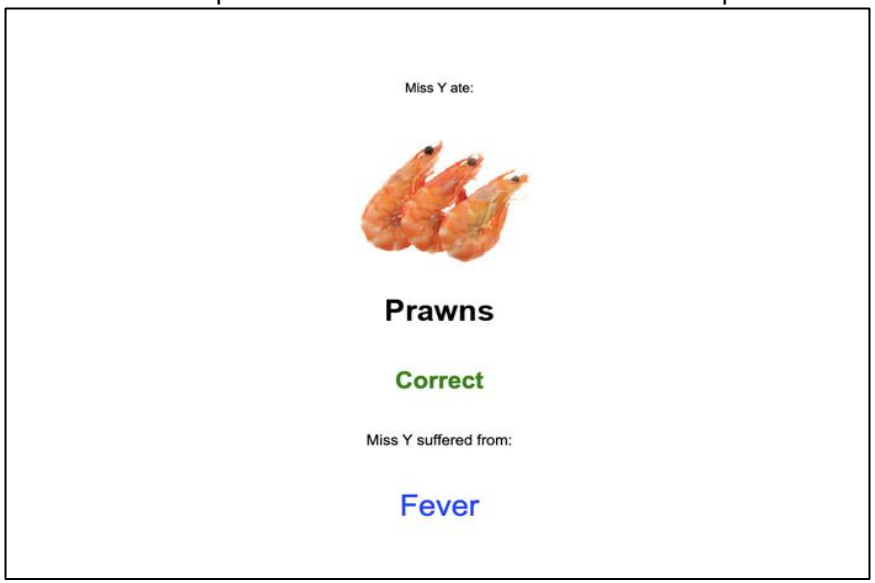
Example training trial




Example of corrective feedback for an incorrect response




Example of corrective feedback for a correct response



Example of a memory test trial

Example of a causal rating test trial

**Experiment 4.4**

Example of a training trial from the passive learning mode condition


# Appendix C: Means tables

## Chapter 2

### Experiment 2.1

| | *Learning score* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (E-H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler non-allergenic (M, N, U, X-Z) | Filler allergenic (V & W) |
| Same patient | 80.72 (6.64) | 38.64 (7.23) | 57.08 (6.47) | -21.29 (5.95) | 68.11 (7.36) | 81.54 (4.69) | 76.98 (7.27) |
| New patient | 78.02 (5.53) | 57.01 (7.79) | 68.59 (5.09) | 22.46 (8.95) | 68.39 (6.01) | 80.27 (4.55) | 84.87 (4.63) |
| | *Causal rating* | | | | | | |
| Same patient | 96.65 (1.63) | 53.12 (4.46) | 65.94 (3.15) | 24.63 (4.99) | 83.03 (3.25) | 7.48 (1.87) | 95.90 (1.79) |
| New patient | 83.67 (3.07) | 73.68 (4.28) | 75.04 (3.35) | 60.40 (5.54) | 75.57 (4.02) | 19.46 (4.42) | 92.81 (2.58) |

### Experiment 2.2

| | *Learning score* | | | | | |
|---|---|---|---|---|---|---|
| | Pretrained blocking (A, C, & E) | Blocked (B, D, &F) | Overshadowed (G-L) | Pretrained unovershadowing (M, O, & Q) | Unovershadowed (N, P, & R) | Filler non-allergenic (Y, Z, a-d) |
| Same patient | 85.26 (5.07) | 36.48 (6.71) | 55.81 (6.19) | -4.61 (4.83) | 68.19 (5.95) | 88.47 (3.54) |
| New patient | 73.80 (7.50) | 37.79 (5.78) | 56.91 (5.73) | 23.45 (7.16) | 69.71 (5.94) | 84.23 (3.93) |
| | *Causal rating* | | | | | |
| Same patient | 92.05 (2.33) | 48.79 (4.01) | 65.48 (3.06) | 19.79 (3.62) | 84.47 (2.95) | 9.17 (2.55) |
| New patient | 88.29 (2.62) | 70.95 (4.69) | 77.53 (2.99) | 51.90 (5.93) | 82.45 (3.29) | 13.30 (2.93) |

**Experiment 2.3**

| 2.3a | *Learning score* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (E-H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler non-allergenic (M, N, U, X-Z, a, b) | Filler allergenic (V & W) |
| Same patient | 89.17 (2.44) | 55.23 (4.53) | 66.48 (3.59) | -9.01 (4.02) | 79.43 (3.60) | 85.71 (2.94) | 88.64 (3.16) |
| New patient | 79.46 (3.18) | 51.79 (4.75) | 69.97 (3.20) | 39.31 (5.79) | 62.44 (4.30) | 82.22 (3.26) | 84.19 (3.10) |
| | *Causal rating* | | | | | | |
| Same patient | 93.85 (1.31) | 56.14 (2.62) | 68.82 (1.76) | 19.59 (2.76) | 86.00 (1.81) | 5.73 (1.06) | 96.27 (0.94) |
| New patient | 80.23 (1.95) | 68.53 (1.76) | 73.49 (1.75) | 60.81 (3.24) | 73.07 (2.18) | 6.43 (1.07) | 92.02 (1.30) |

| 2.3b | *Learning score* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (E-H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler non-allergenic (M, N, U, X-Z, a, b) | Filler allergenic (V & W) |
| Same patient | 84.80 (2.68) | 51.21 (4.52) | 65.04 (3.46) | -8.14 (3.92) | 72.79 (3.80) | 84.82 (2.39) | 88.74 (2.20) |
| New patient | 77.31 (3.78) | 64.97 (4.96) | 70.82 (3.25) | 47.82 (5.48) | 71.97 (4.82) | 87.81 (2.39) | 91.55 (2.13) |
| | *Causal rating* | | | | | | |
| Same patient | 93.38 (1.34) | 57.87 (2.41) | 69.48 (1.92) | 26.97 (3.06) | 85.95 (2.01) | 7.20 (1.28) | 96.46 (0.78) |
| New patient | 78.85 (2.41) | 71.87 (2.66 | 74.87 (2.09) | 60.16 (3.24) | 76.39 (2.18) | 6.95 (1.10) | 93.51 (1.42) |

**Chapter 3**

**Experiment 3.1**

| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (instructed cues)* (E & G) | Overshadowed (non-instructed cues)* (F & H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler allergenic (non-instructed)* (M & N) | Filler allergenic (instructed)* (S & T) |
|---|---|---|---|---|---|---|---|---|
| *Learning score* | | | | | | | | |
| New patient Neutral | 70.66 (7.36) | 42.25 (9.54) | 78.75 (6.46) | 71.53 (6.30) | 35.24 (12.09) | 46.40 (8.62) | 88.51 (4.56) | 78.71 (6.34) |
| New patient Known Cause | 96.21 (1.03) | 34.62 (7.43) | 93.19 (3.50) | 43.63 (7.78) | 86.81 (5.32) | 30.32 (7.09) | 89.98 (3.72) | 93.49 (3.20) |
| *Causal rating* | | | | | | | | |
| New patient Neutral | 66.75 (3.88) | 50.96 (4.17) | 60.40 (3.47) | 66.57 (3.53) | 65.29 (4.53) | 61.04 (3.67) | 82.24 (3.97) | 85.72 (3.13) |
| New patient Known Cause | 89.72 (2.98) | 23.91 (3.62) | 86.70 (3.60) | 30.13 (4.16) | 83.96 (4.21) | 23.07 (3.63) | 80.05 (4.74) | 90.48 (3.15) |

*Note.* Means reported here include the data from the outlier identified in the known cause group. Categories marked with an * are only relevant to the known cause group who received explicit instructions detailing which cues were allergenic.

**Experiment 3.2**

| | Pretrained blocking (A & C) | Blocked (B & D) | Instructed Overshadowed (instructed cues)* (E & G) | Instructed Overshadowed (non-instructed cues)* (F & H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Non-instructed Overshadowed (M-P) | Filler allergenic (Q & R) | Filler non-allergenic (S & T) |
|---|---|---|---|---|---|---|---|---|---|
| *Learning score* | | | | | | | | | |
| New patient Neutral | 68.24 (7.02) | 34.70 (8.33) | 61.97 (8.25) | 56.17 (8.56) | 45.52 (9.27) | 44.16 (7.81) | 52.60 (7.59) | 78.63 (7.36) | 63.95 (7.14) |
| New patient Known Cause | 87.59 (6.20) | 15.83 (8.80) | 88.31 (4.36) | 19.98 (9.07) | 80.34 (6.60) | 17.31 (8.07) | 43.78 (8.02) | 84.99 (6.34) | 70.49 (7.48) |
| *Causal rating* | | | | | | | | | |
| New patient Neutral | 76.56 (3.25) | 63.74 (2.98) | 75.16 (2.64) | 67.72 (2.89) | 64.84 (4.24) | 63.67 (3.45) | 67.32 (2.63) | 90.77 (1.94) | 19.85 (4.83) |
| New patient Known Cause | 95.16 (1.21) | 38.36 (4.69) | 94.86 (1.11) | 40.54 (4.44) | 89.20 (2.75) | 40.54 (4.23) | 62.78 (2.79) | 93.67 (2.12) | 9.90 (3.15) |

*Note.* Categories marked with an * are only relevant to the known cause group who received explicit instructions detailing which cues were allergenic.

## Chapter 4

### Experiment 4.1

| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (E-H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler non-allergenic (M, N, U, X-Z) | Filler allergenic (V & W) |
|---|---|---|---|---|---|---|---|
| *Learning score* | | | | | | | |
| New Patient | 69.31 (6.08) | 38.59 (7.92) | 66.33 (6.58) | 36.72 (8.51) | 67.41 (6.86) | 72.88 (7.19) | 81.81 (5.19) |
| Proactive practice | 70.07 (7.48) | 56.96 (7.32) | 70.48 (5.78) | 31.76 (10.06) | 59.28 (7.87) | 87.40 (4.08) | 86.65 (5.47) |
| Retroactive practice | 87.87 (3.67) | 73.32 (7.36) | 79.70 (5.35) | 59.29 (8.13) | 72.88 (5.90) | 89.85 (4.59) | 91.98 (3.64) |
| *Causal rating* | | | | | | | |
| New Patient | 83.44 (2.73) | 75.06 (3.77) | 78.09 (3.19) | 63.14 (4.21) | 77.81 (3.32) | 10.40 (2.65) | 93.68 (1.66) |
| Proactive practice | 81.03 (2.75) | 68.12 (3.72) | 73.94 (3.07) | 60.76 (5.29) | 71.51 (3.79) | 10.83 (2.66) | 90.89 (2.14) |
| Retroactive practice | 84.18 (3.47) | 76.09 (3.93) | 79.58 (3.58) | 69.04 (4.90) | 78.79 (3.54) | 13.60 (3.55) | 97.35 (0.94) |

**Experiment 4.2**

| | *Learning score* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (E-H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler non-allergenic (S – X, a & b) | Filler allergenic (Y & Z) |
| Same Patient | 89.03 (2.57) | 27.97 (5.53) | 56.44 (4.03) | -23.18 (3.87) | 68.12 (4.54) | 86.57 (2.93) | 84.33 (3.63) |
| New Patient | 81.05 (3.29) | 53.24 (4.43) | 68.74 (3.45) | 44.42 (5.66) | 67.83 (4.04) | 82.02 (3.28) | 87.53 (3.11) |
| Same patient + practice | 84.10 (2.85) | 55.00 (4.90) | 67.88 (3.43) | 28.68 (5.47) | 63.35 (4.28) | 83.72 (3.30) | 88.60 (3.08) |
| New patient + practice | 77.58 (3.35) | 63.57 (4.05) | 67.94 (3.51) | 41.19 (5.50) | 63.04 (4.27) | 86.38 (2.22) | 88.58 (2.95) |
| | *Causal rating* | | | | | | |
| Same Patient | 92.96 (1.69) | 46.65 (2.81) | 60.48 (1.96) | 20.52 (2.84) | 81.41 (2.48) | 10.53 (1.78) | 94.49 (1.45) |
| New Patient | 78.19 (2.17) | 63.42 (2.66) | 69.51 (2.46) | 57.69 (3.00) | 72.54 (2.71) | 7.90 (1.61) | 93.18 (1.53) |
| Same patient + practice | 82.73 (2.42) | 59.24 (3.14) | 66.26 (2.37) | 56.21 (3.67) | 68.59 (2.62) | 7.98 (1.65) | 93.40 (1.90) |
| New patient + practice | 79.45 (2.36) | 69.56 (2.65) | 72.23 (2.32) | 56.88 (3.23) | 72.16 (2.70) | 15.03 (2.69) | 91.63 (1.91) |

**Experiment 4.3**

*Train patient test*

| | *Learning score* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (E-H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler non-allergenic (U, X-Z) | Filler allergenic (V & W) |
| Same patient | 80.76 (4.81) | 26.17 (8.93) | 59.24 (6.67) | -25.83 (5.21) | 65.09 (7.51) | 89.96 (3.59) | 84.91 (5.98) |
| New patient | 71.80 (6.22) | 55.68 (7.80) | 68.10 (6.13) | 49.08 (5.85) | 66.87 (6.99) | 78.10 (5.85) | 77.54 (7.60) |
| | *Causal rating* | | | | | | |
| Same patient | 92.09 (2.81) | 48.51 (4.52) | 64.91 (2.97) | 14.74 (4.15) | 85.53 (3.14) | 10.25 (3.93) | 95.07 (2.05) |
| New patient | 82.53 (2.94) | 74.29 (3.88) | 79.05 (3.28) | 69.92 (4.81) | 81.40 (3.43) | 14.04 (3.94) | 90.71 (2.63) |

*Pretrain patient test (new patient group only)*

| | Pretrained blocking (A & C) | Pretrained unovershadowing (I & K) | Filler non-allergenic (M & N) | Filler allergenic (O - T) |
|---|---|---|---|---|
| *Learning score* | 63.58 (7.61) | 76.78 (6.92) | 82.03 (6.36) | 72.17 (5.63) |
| *Causal rating* | 84.76 (3.29) | 12.71 (4.56) | 14.74 (4.58) | 80.46 (2.89) |

**Experiment 4.4**

*Learning scores*

| | *Active* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (E-H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler non-allergenic (S – X, a & b) | Filler allergenic (Y & Z) |
| Same patient | 87.88 (2.66) | 39.61 (5.21) | 58.22 (3.66) | -21.29 (3.89) | 64.19 (5.25) | 84.17 (3.17) | 87.36 (2.64) |
| New patient | 81.33 (4.12) | 55.02 (5.66) | 68.19 (3.95) | 41.04 (6.30) | 67.30 (4.63) | 78.98 (3.73) | 86.87 (3.77) |
| | *Passive* | | | | | | |
| Same patient | 76.41 (3.75) | 40.77 (5.10) | 53.97 (4.68) | -26.55 (3.78) | 62.13 (5.10) | 79.64 (3.43) | 72.12 (4.65) |
| New patient | 74.47 (4.79) | 51.66 (5.35) | 61.17 (4.64) | 32.71 (6.16) | 49.14 (5.84) | 84.05 (3.36) | 75.00 (3.98) |

*Causal Ratings*

| | *Active* | | | | | | |
|---|---|---|---|---|---|---|---|
| | Pretrained blocking (A & C) | Blocked (B & D) | Overshadowed (E-H) | Pretrained unovershadowing (I & K) | Unovershadowed (J & L) | Filler non-allergenic (S – X, a & b) | Filler allergenic (Y & Z) |
| Same patient | 92.89 (1.74) | 53.00 (3.30) | 64.96 (2.37) | 18.49 (3.18) | 84.12 (2.52) | 5.81 (1.20) | 93.81 (1.46) |
| New patient | 79.99 (2.39) | 64.35 (3.17) | 70.61 (2.48) | 55.69 (3.82) | 73.98 (2.55) | 9.35 (1.70) | 95.81 (1.25) |
| | *Passive* | | | | | | |
| Same patient | 91.65 (2.06) | 48.27 (3.00) | 62.24 (2.66) | 12.73 (2.70) | 85.40 (2.38) | 12.10 (2.43) | 91.15 (2.37) |
| New patient | 81.71 (2.36) | 66.86 (3.09) | 69.31 (2.54) | 56.92 (3.92) | 71.12 (3.21) | 9.14 (1.88) | 93.97 (1.62) |

# Appendix D:

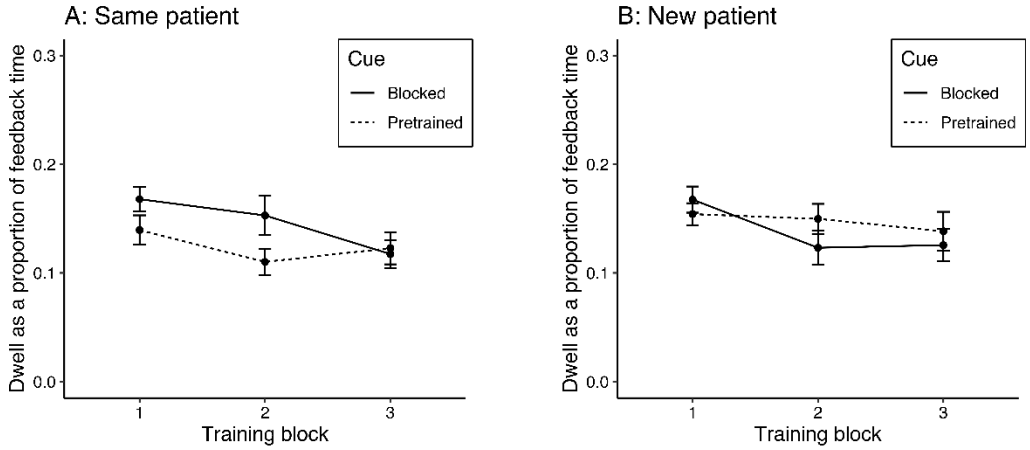
## Analysis of eye-gaze from Experiment 2.2 (Feedback period)

Trials with no recorded fixations for the relevant period were removed from the analyses (*M* = 3.18%, *SE* = 0.53 for the feedback period of training). Any trials with no recorded gaze on either cue were also excluded from the analysis (*M* = 22.71%, *SE* = 1.85 for the feedback period of training). Dwell time was aggregated over cues of the same type, for instance dwell time on the blocked cue is the mean of the proportion dwell time for cues B, D and F.

Dwell time on the critical cues during the feedback period as a proportion of the total feedback time (two seconds) is shown in Figure D1. Overall, there was moderate evidence that participants showed no bias in the proportion of time spent looking at the blocked cue over the pretrained cue during the feedback period, $BF_{01}$ = 3.898, *error%* = 0.808, $\eta_p^2$ = 0.015, $F < 1$. Evidence for the absence of an interaction between cue and instructions was weak, $BF_{Excl}$ = 1.312, $\eta_p^2$ = 0.033, $F(1,55) = 1.882$, $p = .176$. There was a decline in attention to both blocked and pretrained cues during feedback regardless of instruction group, $BF_{10}$ = 14.389, *error%* = 3.107, $\eta_p^2$= 0.110, $F(2,110) = 6.811$, $p = .002$, for the effect of block. The evidence favoured excluding all interactions, $BF_{Excl}$ = 3.306, $\eta_p^2$= 0.034, $F(2,110) = 1.942$, $p = .151$ against the block x cue interaction, $BF_{Excl}$ = 14.430, $\eta_p^2$= 0.005, $F < 1$ against the block x instruction interaction, and $BF_{Excl}$ = 3.596, $\eta_p^2$= 0.023, $F(2,110) = 1.305$, $p = .275$ against the three-way interaction.

**Figure D1**

*Eye-gaze data from the feedback period of training in Experiment 2.2*




*Note*: Eye-gaze during the feedback period of the training phase. Panels A and B show the mean proportion of feedback time spent looking at each cue for the blocking compounds in the Same patient group and New patient group respectively. Error bars represent the standard error of the mean.