The Redundancy Effect in Human Causal Learning

Shu Chen

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

2026

Statement of Originality

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

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

Shu Chen, 30.06.25

Artificial Intelligence

No content produced by generative AI tools has been used in the preparation of this thesis.

Australian Government Support

The research reported in this thesis was supported by the award of an Australian Government Research Training Program (RTP) Scholarship to the PhD Candidate.

Acknowledgement

I would like to express my gratitude to my supervisor, Evan Livesey, for his guidance in the past five years. This thesis would not have been possible without his unwavering support, academic dedication, and technical expertise, for which I am truly grateful. I would also like to thank Justin Harris and Robert Boakes, who reviewed my research proposal and provided constructive feedback on this project.

My appreciation goes to the supportive and collaborative laboratory group. In particular, I would like to thank Jessica Lee, who offered valuable suggestions on empirical chapters; Justine Greenaway, who guided me through the human testing procedure with insight and encouragement; and Dominic Tran, who shared insightful ideas on experimental designs.

Finally, I am thankful of my family, for their care and support throughout this journey and beyond.

Abstract

Everyday life requires us to make judgments based on indirect and fragmental evidence. This is especially true when we acquire causal knowledge, where potential causes are often correlated, redundant, and thus highly ambiguous. The complexity of ambiguous causal learning situations is epitomised by a recent human learning phenomenon known as the redundancy effect. It refers to the comparison between two types of redundant cue: a blocked cue, which predicts a meaningful outcome but supplies no new or unique information about its occurrence, and an uncorrelated cue, which is always less informative about the outcome than the cues with which it coincides. While the redundancy effect represents a novel empirical test of classic theories of associative learning, which appear to be broadly challenged by the different ways that people learn about redundant cues, proposition-based reasoning via higher order cognition could shed light on the causality judgment process and how it varies under different *a priori* assumptions. The current thesis took an integrated approach to understand differential learning about redundant cues in a range of ambiguous conditions. Past literature identified prediction error learning algorithms and attention deployment strategies as promising mathematical models for explaining the redundancy effect. Previous research also suggests that inferring causal relationships for the blocked and uncorrelated cues is influenced by magnitude additivity and preventative cue assumptions. To explore the involvement of simple associative mechanisms, Chapter 2 conducted computational modelling of various candidate models to compare their simulated predictions with empirically observed data. To examine the role of inferential reasoning, Chapter 3 manipulated prior assumptions necessary for deductive inference about redundant cues in a pretraining phase. Chapter 4 examined the influence of deductive reasoning about redundant cues on the subsequent tendency to produce a learning and judgment bias referred to as theory protection. Chapter 5 provides further evidence that the types of learning scenario typically used in causal learning research have a consistent effect on how people engage in causal reasoning about redundant cues, in a way that is critical for understanding the existing literature on the redundancy. The results are consistent with a multi-process explanation for the redundancy effect, suggesting that learning and memory processes driven by prediction error and propositional reasoning are both important determinants in human learning under ambiguity.

Table of Contents

List of Tables

List of Figures

Chapter 1

General Introduction

The capacity to extract complex causal structure from an ever-changing world is a fundamental aspect of human cognition. Causal learning allows us to predict ensuing events, make informed decisions, and modify behaviours to avoid aversive consequences and accomplish desired goals. The study of associative learning lays the foundation of our understanding of how humans cognise environmental signals differently depending on their causal significance. Research within this field has traditionally focused on explaining learning about cues that are predictive of meaningful outcomes. A famous demonstration of association formation between two independent events comes from Pavlov (1927) in his classical conditioning experiments. He presented his dogs with a ticking metronome which was shortly followed by the provision of food. After repeated contiguous sound-food pairings, his dogs learned to associate the ticking sound with food delivery and salivated upon hearing the metronome in anticipation of foods. The Pavlovian conditioning paradigm has since been adapted to study human behaviour in a variety of different predictive learning contexts, including causal learning. Parallels between patterns of anticipatory behaviour in classical conditioning and patterns of explicit judgments in human causal learning have led researchers to suggest that the basic associative processes that manifest in animal learning can mediate the extent to which humans judge one event as being causal of the other (Dickinson, 2001). A wealth of causal learning research has provided insights into how we draw cause-and-effect relationships from past experiences and apply causal knowledge in future situations that predict reward (e.g. Seymour et al., 2005; van den Akker et al., 2008), pose threat (e.g. Hygge & Ohman, 1978; Olsson & Phelps, 2004), and of most relevance to the current research —— require causality judgment to solve problems (e.g. Dickinson et al., 1984; Dickinson & Burke, 1996).

Not only is it important to understand how we learn causal connections between informative cues and the outcomes that they predict, but it is equally important to understand how we learn, or *fail* to learn, about cues that do not convey useful information. Imagine in a situation where a patient takes medication daily to regulate hormone levels, the patient tests a variety of medications to find the one that best stabilises their condition. If medicine A repeatedly leads to a hormone change when taken on its own, the patient would establish a causal relationship between medicine A and hormone change. If, at a later point, the patient takes medicine A together with medicine X and experiences the same change, even though the effect of X has not been seen in isolation, the patient might regard X as not adding anything important to the effect of A and hence non-causal of hormone change. In this instance, competition between contemporaneous cues A and X for the strength of causal relationship with the outcome might lead to judgment of X as non-causal. This is one instantiation of ambiguous causal learning where the effect of potential causes is not directly observable but needs to be indirectly inferred from fragmentary evidence.

While it is common for the presence of other competing causes to modulate learning in a way that affects how the causal relationship between a cue and an outcome is judged, not all redundant cues are equally disregarded as potential causes. Indeed, recent research comparing the fate of ambiguous cues from different cue competition paradigms has revealed interesting disparities in causal judgment (Uengoer et al., 2013; Uengoer et al. 2020), perceived confidence (Jones et al., 2019), subsequent learning (Spicer et al., 2020), as well as potential variations in attention deployment (Jones & Zaksaite, 2018; Uengoer et al., 2019) and prior assumptions (Jones et al., 2019; Zaksaite & Jones, 2020). These studies consistently suggest an overarching role for ambiguity in the interpretation of causality among complex events in the environment. With a view to building further upon existing knowledge about human learning about redundant cues, the current thesis will explore a range of ambiguous

learning situations centered around a phenomenon known as the *redundancy effect*, which involves a comparison of learning about different redundant cues (Pearce et al., 2012; Uengoer et al., 2013). I will first describe the redundancy effect and explain its theoretical significance, before introducing the impact of uncertainty and deductive reasoning in causal learning.

Table 1.1

Key design elements and findings in the blocking, relative validity and redundancy effects

Training	Typical findings
Blocking: A+ AX+ DE+	Blocking effect: X < E
Relative Validity: BY– CY+ FG+/– HG+/–	Relative Validity effect: Y < G
	Redundancy effect: X > Y

The Redundancy Effect

Tables 1.1 presents a summary of the critical findings discussed in this section. A classic example of cue competition is provided by the phenomenon of blocking (Kamin, 1969). In a blocking procedure (e.g. A+, AX+), pairing a compound holding two elements A and X with an outcome results in strong associative strength to A and reduced associative strength to X if A has previously been established as a strong predictor of the outcome on its own. It should be noted that in a conventional blocking design, the A+ trials are delivered before the AX+ trials in a phased manner, whereas in the investigation of the redundancy effect, the two trials types are typically intermixed. Blocking is said to be evident if responding to X is less than responding to D or E from a similar two-element compound where neither constituent cue has been individually trained with the outcome (e.g. DE+). While D and E retard learning about each other compared to when they are trained as single

cues (i.e. the overshadowing effect; Mackintosh,1976), they are nonetheless seen as more plausible causes than X. This observation of blocking (D/E>X) suggests that mere co-presence with the outcome does not guarantee learning for a cue but the predictive value of accompanying cues must also be considered. The usefulness of X is discounted by the fact that it conveys no additional, but rather the exact same, information as A. The blocked cue X from the blocking design can thus be described as *informationally redundant*.

Likewise, the relative validity effect (Wagner, 1968) refers to the finding that learning about the common cue Y from relative validity (e.g. BY–, CY+) is considerably poorer than learning about its counterpart G from pseudo-discrimination (e.g. FG+/–, HG+/–). In this case, both common cues Y and G are trained under the same 50% partial reinforcement schedule yet they possess different predictive power relative to their associates. The prediction that the outcome would occur 50% of the time makes Y an unreliable signal of outcome non-occurrence compared to B which never leads to the outcome and an unreliable signal of outcome occurrence compared to C which always leads to the outcome. Although G is also paired with the outcome on half of the occasions, it is just as good a predictor as its accompanying F and H as all three cues predict the outcome with a probability of 0.5. The low predictive validity of Y relative to the reliable predictors that co-occur with it hinders the formation of causal relationship between Y and the outcome. This disadvantage is no longer present when the predictive utility is equalised across all concomitant cues. The uncorrelated cue Y from a relative validity can thus be described as *predictively redundant*.

When it is observed in human causal learning, the redundancy effect refers to the observation that causal judgments about the blocked cue X from the blocking treatment are consistently higher than those for the uncorrelated cue Y from the relative validity preparation. The finding was first documented in Pavlovian conditioning in laboratory animals (Pearce et al., 2012) but has been studied more extensively in human causal learning

(Jones & Zaksaite, 2018; Jones et al., 2019; Uengoer et al., 2013; Uengoer et al., 2019; Uengoer et al., 2020; Zaksaite & Jones, 2017; Zaksaite & Jones, 2020). The redundancy effect has sparked an increasing interest into the ultimate fate of ambiguous cues in learning due both to its theoretical significance and practical implications. As will be reviewed in depth in later sections, the finding of unequal learning about the blocked and uncorrelated cues represents a major challenge to traditional associative analyses (e.g. the Rescorla-Wagner model; Rescorla & Wagner, 1972; Wagner & Rescorla, 1972) as none could provide a satisfactory account. In the context of real-world applications, investigating and comparing redundant information from different sources is at the heart of understanding how and why people form causal beliefs under certain circumstances but not under others. These basic learning processes also serve as a valuable tool for gaining insight into the underlying mechanisms of many clinical conditions. For example, a deficit in selective formation of causal links has been suggested as one of the mechanisms giving rise to non-specific fears in anxious individuals (Boddez et al., 2012) and both positive and negative symptoms in patients with schizophrenia (Moran et al., 2008; Morris et al., 2013).

Absolute Relationship and Relative Informativeness

While both the blocked and uncorrelated cues are considered redundant in their respective designs, two intuitive explanations for why they are treated differently have been speculated from the nature of training that they receive: one in terms of absolute relationship and the other in terms of relative informativeness. Pearce and colleagues (2012) surmised that the blocked cue (henceforth X) accrues higher associative strength than the uncorrelated cue (henceforth Y) because X is continuously but Y is partially paired with the outcome. The reinforcement schedule under which the redundancy is established may contribute to greater learning about the cue that more likely results in the outcome. The authors further noted that the two redundant cues also differ with respect to their informational value relative to

accompanying cues. X is equally informative about outcome presence as its associate A but Y is less correlated with either the presence or the absence of the outcome than its associates B or C. The higher relative informativeness for X than for Y may result in the former eliciting a greater outcome expectation than the latter. Both of these theoretical constructs, absolute relationship and relative informativeness, suggest that even though X provides no new information above and beyond that of A, it may nonetheless retain a higher functional value than Y for its higher outcome probability and relative informational utility.

Uengoer and colleagues (2013) proposed that absolute relationship and relative informativeness are distinctive enough to be treated as separate theoretical constructs. They hypothesised that retrieval of memories of the outcomes that occurred with the cues may be key to the causal ratings that participants make. They noted that cue Y would retrieve representations of both outcome presence and outcome absence, whereas cue X would only bring to mind the representation of outcome presence. This difference in memory retrieval (with conflicting memories present for Y but not X) provides one possible psychological explanation for why X is regarded as a stronger cause than Y, and one specifically based on the absolute relationship between the cue and outcome. They argued that if this were the case, then changing the cue-outcome contingencies—balancing outcome probability for X and Y while maintaining a difference between them in terms of their relative informativeness—could eliminate any difference in retrieval conflict and hence the redundancy effect. That is, equating the absolute relationship between both redundant cues and the outcome should allow the sole influence of relative informativeness to be observed. To disentangle the two constructs, the authors invoked a 50% partial reinforcement schedule for the blocking procedure (both A alone and AX in combination resulted in outcome presence 50% of the time and outcome absence 50% of the time). This ensured that both redundant cues established the same absolute relationship with the outcome independent of the usefulness of

the cues accompanying them. Even when the blocking contingencies were partially reinforced in this fashion, Uengoer et al. found that causal ratings for X were higher than causal ratings for Y. This suggests that the higher relative informativeness of X compared to Y is sufficient to drive the redundancy effect. Although these explanations in terms of the statistical properties of the contingencies are admittedly informal, they may provide a simple heuristic approach to the redundancy effect in situations where other formal accounts fare less well. The plausibility of these theoretical constructs will be evaluated in relation to formalised mathematical learning algorithms in Chapter 2.

Associationist Approach to the Redundancy effect

Upon the initial discovery of the redundancy effect, researchers have attempted to apply traditional associative analyses to locate the mechanisms responsible for the differences in outcome expectation that the blocked and uncorrelated cues evoke when assessed independently. The following section will introduce associative theories proposed in the literature as candidate explanations for the redundancy effect. In particular, I will take the popular approach of implementing a prediction error learning algorithm (e.g. the Rescorla-Wagner model) and examine the effect of varying aspects of its specification (e.g. as explained below, providing a common element that is shared among cues, using a summed vs. separable error term, implementing changes in attention). These models will constitute the focus of computational analyses in Chapter 2. Explanations that rely on within-compound associations have contributed to a significant part of the research work but bear less relevance to the current thesis. These explanations will be described in full detail in this section and mentioned only briefly in later chapters.

The most prevailing accounts from an associationism tradition are based on error-correction learning algorithms. The Rescorla-Wagner (1972) model is a foundational example of such where the formation of an association between a cue and an outcome is

thought to proceed only to the extent that the outcome is surprising, and thus slows as the outcome becomes accurately predicted. The model posits that learning is governed by the discrepancy between the prediction derived from all cues available at the same time and the actual outcome realised (i.e. a summed error term). Applying this rule to the blocking design, the blocking cue A is expected to reach near-asymptote levels of learning after being separately reinforced on its own. When presented together with the blocked cue X on AX trials, A reliably predicts the outcome rendering little prediction error to fuel further learning about X. The model thus predicts that X should elicit very weak outcome expectation on test. Turning to relative validity, the uncorrelated cue Y is expected to gain associative strength on reinforced CY trials and lose associative strength on non-reinforced BY trials. The loss is however of a smaller magnitude than the gain by virtue of the companion B gradually developing negative associative strength. That is, the inhibitory strength of B allows some of the excitatory strength of Y accrued on CY trials to be retained. The model thus predicts that, on test, Y should elicit a moderate expectation of the outcome. The pattern of predictions generated by the Rescorla-Wagner model and a broad class of other summed error models (e.g. Pearce, 1987; Gluck & Bower, 1988; Esber & Haselgrove, 2011), namely that anticipation of the outcome should be stronger for Y than for X, is at odds with the empirically observed learning advantage for the blocked cue in the redundancy effect (Pearce et al., 2012).

The theoretical challenge posed to the Rescorla-Wagner model has led Vogel and Wagner (2017) to propose a modification to the original formulation to encompass the influence of some common attributes that are assumed to be shared among all cues. Specifically, the modified Rescorla-Wagner model assumes that a common cue, denoted by K, may be present on all training and testing trials. Because this common component is activated on all trials, X and Y are competing with AK and CK on respective reinforced

trials. For X, the pattern of associative change remains almost identical to the predictions of the original model with very little learning for X from a staged blocking design or with initial increments followed by decrements of equal magnitude for X from an intermixed blocking design. For Y, however, these predictions become very different. Y not only suffers a greater loss of associative strength on nonreinforced BYK trials due to overexpectation of the outcome (i.e. summed associative strength for B, Y, and K exceeds asymptotic strength supported by the outcome), it also acquires less associative strength on reinforced CYK as learning is blocked on those trials by the common K. Consequently, Y undergoes a larger total associative decrement than X and elicits a weaker outcome expectation than X at any point during training (the same pattern of results will be observed at test because K is common to both cues). Hence, the Rescorla-Wanger model with a common element assumption is able to predict the redundancy effect.

Although the common element model has achieved considerable success in explaining the key findings of the standard redundancy effect, it struggles to reconcile with related results from several studies. One of the empirically testable predictions from the common element approach is that when the proportion of reinforced trials is low, the influence of the common cue on Y will be weak. As such, when X and Y are presented at test, they should essentially elicit the same outcome expectation. Jones et al. (2019) investigated the effect of outcome base rate on the redundancy effect by varying the percentage of trials that were followed by the outcome. The authors found a weaker yet marginally significant redundancy effect when the overall outcome rate was reduced from 75% to 25%. Moreover, Uengoer et al. (2020) trained participants with A alone trials prior to AX compound trials in a staged manner instead of intermixing the blocking contingencies. This change in methodology for observing blocking, according to Vogel and Wagner (2017), should lead to similar associative strengths for X and Y. The fact that Uengoer and colleagues continued to

demonstrate stronger causal ratings for the former suggests that the common element approach may not be well-suited as an adequate explanation for the redundancy effect.

While summed error models do not appear to accommodate the full results of the redundancy effect, the basic error-correction algorithm may remain plausible if it is asserted that prediction error is calculated on the basis of mismatch between each individual cue and the outcome (i.e. individual error term). Individual error models enable cues to reach the asymptote of learning supported by the outcome regardless of the associative value of other simultaneously presented cues but relies on additional processes such as attention to capture competition between cues. The most influential example is perhaps the theory of selective attention proposed by Mackintosh (1975) which combines individual error learning algorithm with a preferential attention mechanism that favours relatively good predictors. More recently, Uengoer et al. (2020) updated Mackintosh's theory by proposing a variation to the rules governing attention during associative learning. Central to both of these attentional accounts is the assumption that associability (i.e. the readiness of a cue to enter into an association with the outcome) is not an immutable property, but changes depending on a cue's ability to predict an outcome of significance. Specifically, the associability of a cue rises if it predicts an outcome that is otherwise unpredictable and falls if it is of low predictive utility relative to concomitant others. It follows from this assumption that attention will be prioritised for cues possessing higher predictive validity at a cost to the processing of less useful competitors. In the redundancy effect, the blocking cue A from the blocking procedure and the distinctive cues B and C from the relative validity design are better able to predict their respective outcomes than their associates and will thus attract more attentional resources. Consequently, cues A, B, and C will be more strongly associated with their respective outcomes than the accompanying cues X and Y, leading to the blocking and relative validity effects. Although attentional accounts do not make specific predictions about

the rate of associability change for different redundant cues, learning differences may be explainable by assuming different declining rates. That is, if one conceives of the possibility that the associability of Y approaches zero faster than does the associability of X and Y has not accrued enough associative strength before becoming non-associable, then attention accounts can naturally predict less associative strength for Y than for X. Simulated associability differences will be presented in more detail in Chapter 2.

Attentional theories like those proposed by Mackintosh (1975) and Uengoer et al. (2020) conceptualise cue competition as being the product of learning to attend to cues with more predictive power and ignore those that have less. As such, their veridicality is typically assessed by studying the attention paid to the cues, using a range of overt and covert measures. While overt attention is often assessed by measuring eye gaze (greater dwell time on a cue is assumed to be synonymous with greater attention), covert attention is often gauged by measuring changes in associability, in particular the speed with which cues enter into new associations. Attention has been examined in a number of blocking studies (e.g. Beesley et al., 2011; Kruschke et al., 2005; Kruschke & Blair, 2000; Le Pelley et al., 2007; Luque et al., 2018) and discrimination learning studies involving designs that are functionally similar to the relative validity used to study relative validity. For example, Kruschke and Blair (2000) tracked eye gaze in human participants during a blocking procedure. They observed an elevated attention to the blocking cue and a reduced attention to the blocked cue and this attentional difference was found to covary with the magnitude of blocking. Given previous demonstrations of a strong correlation between attention and learning in cue competition effects, one may expect the redundancy effect to similarly reflect an attention difference. To date, two papers have inspected attention using a redundancy effect design (Jones & Zaksaite, 2018; Uengoer et al., 2019). Surprisingly, results from both of these studies point to the conclusion that, between the two redundant cues of interest, there is no

difference in attention either in eye gaze or associability, suggesting that attention may not suffice as a primary account for the greater outcome expectation elicited by the blocked cue over the uncorrelated cue.

Aside from the well-formulated theories described above, accounts based on within compound associations as following naturally from standard associative principles offer an alternative explanation for the redundancy effect. According to a within-compound associative account, presentation of one cue at test leads to retrieval of companion cues via their within-compound associations, which in turn leads to retrieval of the outcomes with which these companion cues are associated. This retrieval serves to mediate the associative strength for a given cue in an additive manner (e.g. Rescorla & Durlach, 1981). In other words, when a cue is presented individually at test, the direction association between the cue and the outcome and the indirect association between its companions and their outcomes sum to yield a final prediction. Within-compound associations predict that being accompanied by a strong predictor A would enhance the ability of X to activate the representation of the outcome. On the other hand, because B and C predict differing outcomes, retrieval of each of these indirect associations would counteract the influence of the other, resulting in a negligible change in the causal judgment for Y. Even though summed error models like that of Rescorla and Wager (1972) predict greater associative strength for Y than for X, X may take additional advantage of within-compound associations to produce a greater outcome expectation than Y. However, evidence that contradicts this claim has been revealed in a number of retrospective revaluation studies. Using an appetitive conditioning procedure for rats and an autoshaping procedure for pigeons, Pearce et al. (2012) devalued the excitatory associate A for X while upvaluing the inhibitory associate B for Y post the redundancy effect training and a test. According to the within-compound association account, revaluation of these companion cues is thought to reduce the strength of the original association between X

and the outcome but enhance the associative strength for Y by indirectly retrieving the updated outcomes associated with the companions. The authors showed that X continued to possess greater associative strength than Y despite successful revaluation of A and B. This result suggests that within-compound associations cannot be the main contributor for differential learning about redundant cues. In a human causality judgment task, Uengeor et al. (2013) similarly carried out revaluation of A and B in the direction opposite in sign to their previous causal status in the blocking and relative validity preparation, respectively. These authors again found a greater tendency for human learners to judge X as a more likely cause than Y, casting further doubt on the adequacy of explanations based on within-compound associations.

The failures confronted by acquisition-focused theories have prompted researchers to consider the redundancy effect as being the result of performance differences at the time of testing. While relying on a similar indirect learning mechanism, the comparator theory makes quite divergent predictions compared to within-compound associations. The comparator hypothesis and its elaborated version (Denniston et al., 2001; Murphy et al., 2001a; Murphy et al., 2001b) propound that associative strength for a cue follows the simple error-reduction rule based on its individually calculated discrepancy with the outcome, but the expression of learned associations is determined through a contrast mechanism. Specifically, the comparator theory stipulates that the activation of outcome representation by a target cue at test is proportional to the strength of direct association between the target and the outcome, inversely proportional to the product of the strength of within-compound association between the target and other co-occurring cues (i.e. first order comparison cues) and the strength of associations between the comparison cues and the outcome. Note that the original comparator hypothesis allows only the companion cue with the strongest association to the target to act as the single dominant comparison cue but its extension allows not only multiple first order

comparison cues to modulate the target's performance but also companions of the first order comparison cues (i.e. higher order comparison cues) to mediate the influence of the first order comparators. For simplicity, the following predictions will focus on first order comparators for the redundant cues X and Y.

In the case of blocking, although X may accumulate high associative strength through consistent pairings with the outcome, comparison with the accompanying A which more consistently evokes an expectation of the outcome would down-modulate the extent to which the X-outcome association is expressed at test. Conversely, activation of the representation of X, which less reliably predicts the outcome, would have less of an impact on performance of A, thus generating blocking. This prediction holds true as long as X accrues lower associative value than A during training. Relative validity can be similarly predicted via a comparator process. Y is involved in trials that terminate with conflicting outcomes, presentation of Y at test would activate representations of comparators B and C which in turn would activate representations of the outcome and its absence. Comparison with its associates would result in Y being a poorer predictor of outcome absence than B and a poorer predictor of outcome presence than C. Although both redundant cues are subject to a down-regulating influence from their companions, the intermittently reinforced Y would have established a weaker direct association with the outcome than the consistently reinforced X before the expression of their learned associations are mediated at test. Assuming similar levels of performance suppression, X would be judged as a more likely cause than Y.

Despite the theoretical plausibility of the comparator theory, the main within-compound contrast mechanism relies heavily on the values of the associated comparator cues, as does the within-compound association account described above. The finding that outcome expectation for X and Y is not influenced by the associative strengths of their associates A and B or vice versa suggests fundamental inadequacy of within compound associations to

explain the redundancy effect (Jones & Pearce, 2015; Pearce et al., 2012; Uengoer et al., 2013), regardless of whether paired cues contribute to outcome expectancies in an additive manner, as in within compound associations, or a subtractive manner, as in comparator theories. In addition, Zaksaite and Jones (2017) tested specific predictions from the comparator hypothesis by focusing on the comparison between X and the consistently reinforced cue C[1] from their relative validity. These authors reversed the significance of A and Y with further A– and Y+ training following the standard redundancy effect procedure. Since A is a first order comparator for X and Y is a first order comparator for C, devaluing A while upvaluing Y should increase the judgement of X being causal but decrease the judgement of C being causal. However, judgement difference between X and C was the same before or after revaluation, suggesting that retrieval of associates at test does not affect causal judgment in the redundancy effect.

Propositional Approach to the Redundancy Effect

Notwithstanding the historical success of the associationist approach in accommodating a wide range of cue competition phenomena across various species, the challenge presented by the redundancy effect to traditional mechanistic models invites reflection on the critical distinctions in the way that animals and humans learn about ambiguous causes. Particularly, humans are capable of generating abstract and transferrable assumptions and beliefs from their past interactions with the world, which they may draw on to construct hypotheses about current ambiguous causal events. In this respect, conceiving cue competition as the product of basic hardwired processes from the perspective of associative analysis may be too simple to capture sensitivity of human causal learning to prior knowledge. Advocates of an approach sometimes referred to as *propositionalism* have instead construed learning as a form of controlled and effortful processing, involving the formulation and evaluation of hypotheses (propositions) about the connectedness of observed

[1]Note that Zaksaite and Jones referred to the reliable predictor of no outcome in the simple discrimination as C rather than B but I have swapped these labels here to keep the nomenclature consistent.

events. The processes involved thus include cognitively demanding analytical and reasoning skills (see Mitchell et al., 2009 for a review). The premise of this account is that people arrive at reasoned decisions about ambiguous cues based on their observations and relevant assumptions about cause and effect. In this thesis, two sets of assumptions are of major interest: they are *magnitude additivity* assumptions for the blocked cue X and *preventative cue* assumptions for the uncorrelated cue Y.

Considerable research on causal inference has shown that judgment of the blocked cue is influenced by whether the effects of multiples cues are perceived as additive (e.g. Beckers et al., 2005; Livesey et al., 2019; Lovibond, 2003). The additive assumption specifies that the effects arising from simultaneous causes should sum to a combined effect of larger magnitude or higher intensity. In the blocking procedure, learners have clear evidence that A independently causes the outcome. When they encounter AX compound, the additive assumption may lead them to expect that, if X is also causal, the combined effect of A and X should be greater than that produced by either alone. The observation that the effect of AX combined is no larger than the normal effect of A then falsifies the causal significance of X. The deductive reasoning process of this kind bears resemblance to the *modus tollens* argument: If X is an effective cause then AX should result in a larger outcome than A alone (if p then q), AX does not result in a larger outcome (not q), therefore X does not cause the outcome (therefore not p).

Given that additive assumptions enable the deductively valid inference that X is a non-causal cue, contradicting such assumptions should improve the plausibility of the alternative. There have been two ways commonly employed in the literature to constrain deductive reasoning based on the default additive assumption. The first is to directly oppose causal additivity by showing that outcomes do not add up to produce a greater effect (Beckers et al., 2005; Livesey & Boakes, 2004; Livesey et al., 2019; Lovibond, et al., 2003; Mitchell et al.,

2005). According to non-additive assumptions, multiple cues together generate the exact same effect as does each individual cue on its own. Under causal non-additivity, the lack of increase in outcome magnitude following AX cannot be taken as clear evidence for X being non-causal because both a causal and a non-causal X are consistent consequents of the same non-additive antecedent. That is, it is possible that X is indeed an effective cause but its causal influence is not observed after applying non-additive rules. The second way to restrict the formation of deductive inference about X is to impose a ceiling on the outcome by showing that A alone has caused the outcome to its maximal extent (De Houwer et al., 2002; Beckers et al., 2005). Even after entertaining additive assumptions, the ceiling effect does not permit outcome additivity on AX trials to be appropriately verified. The lack of increase in the combined effect can be interpreted as either the result of both A and X being causal but there is no opportunity to observe the further contribution of X on the outcome or only A is causal. Outcome non-additivity and maximality both induce ambiguity in the number of possible causes on AX trials, that is, the number of possible causal states that could lead to the same observed effect.

In a similar vein, the relative validity effect involving the uncorrelated cue Y is susceptible to deductive reasoning depending on whether B is considered capable of preventing an outcome from occurring. Y is established as a possible cause when followed by the outcome on CY+ trials but the causal relationship between Y and the outcome is called into question when the outcome is omitted on BY– trials. However, the causal relationship that Y established with the outcome would not be completely discarded if the absence of the outcome on BY– trials is attributed to B mitigating any effects of Y. That is, if B acquires some preventative properties, the outcome that would otherwise follow Y might be assumed to be prevented by B. Zaksaite and Jones (2020)[2] tested whether regarding B as a preventative cue might protect causal beliefs about Y. They gauged the extent to which B

[2] Note that Zaksaite and Jones (2020) described preventative cue-outcome relationships in terms of conditioned inhibition which permits B to develop negative associative strength.

was regarded as preventative during non-reinforced training and found a positive correlation between the extent to which B developed into a preventer and the extent to which Y was judged to be causal.

Zaksaite and Jones (2020) also made an important observation in relation to the degree of preventative influence from B under two different task scenarios, namely, the food allergist task and the hormone change task. In the most conventionally used food allergist task, participants assume the role of a doctor attempting to ascertain which foods are causing a patient's allergies based on the types of foods that the patient has consumed and the allergic reactions that occur or do not occur as a consequence. The food allergist paradigm was compared with the hormone change paradigm in which participants play the role of a medical researcher investigating which medicines are effective in changing hormone levels based on patient records of hormone increase, decrease, or no change following administration of various kinds of medicines. Tests for preventative influence (i.e. presenting a causal cue alongside B) indicated that B became weakly preventative in the hormone change task but no such evidence for preventative cue-outcome relationships was found in the food allergist task. I speculate that this difference arises because it is more common for medicines to prevent physiological (including hormonal) changes than for foods to prevent allergic reactions in real-world situations. Although the redundancy effect has been obtained in both paradigms, the hormone change task may be preferred over the classic allergist task for two reasons. First, the hormone change task reduces the unintended encouragement of deductive reasoning for the uncorrelated cue. Second, to investigate the influence of deductive reasoning, whether to encourage or discourage its use, a task that allows easy manipulation of assumptions is required. The food allergist task is not suited for this purpose as foods are generally considered non-preventative.

There have also been sporadic instances of non-human animals exhibiting signs of complex cognitive functioning akin to causal reasoning in humans (e.g. Bates et al., 2008; Hanus & Call, 2011; O'Connel & Dunbar, 2005). For example, Beckers et al. (2006) showed similar sensitivity in rats to pretrained magnitude additivity rules and outcome ceiling effects using a conditioned suppression protocol (but see Haselgrove, 2010 for a critique). It is, however, beyond the scope of the current thesis to elucidate the parallel processes of inferential reasoning in animal demonstrations of the redundancy effect. There are likely some fundamental associative and memory processes that underlie the development of the redundancy effect across species, but I argue that the symbolic reasoning processes are particularly relevant in the context of human causal learning, firstly because it is known that people possess the capacity for such processes, and secondly because causal learning tasks are explicit cognitive tasks that openly invite the application of such processes.

Dissociative Causal judgment and Confidence Appraisal

Another important finding that emerges from the literature is the dissociation between causal judgment and prediction certainty. Jones and colleagues (2019) asked their participants to indicate the likelihood that X and Y would produce the outcome as well as to report their confidence in these likelihood judgments. X was found to be judged as a more likely cause than Y, demonstrating the redundancy effect, however, certainty with which judgment about X was made was intriguingly lower than that for Y. The inversely correlated ratings for likelihood and confidence judgments prompted the authors to propose a quite distinct view on the redundancy effect in terms of the inherent nature of the task scenario. They argued that, instead of reflecting a difference in the acquisition or the expression of associations formed with the outcome, the redundancy effect reflects a difference in causal ambiguity between X and Y under the food allergy task commonly adopted by human causal judgment experiments. In particular, the food allergy scenario carries with it the assumption

that if certain food causes an allergic reaction, it will elicit the ailment each time it is eaten. Additionally, prevention of allergic reactions by the simultaneous consumption of other foods is generally considered implausible given the rarity of such instances in real-world situations. In this regard, learners can be confident that Y is not a valid cause because it does not always lead to the outcome and outcome prevention is unlikely to occur on trials where it fails to produce the outcome. The causal status of X is less affected by non-preventative assumptions naturally enforced by the food allergy task and thus remains ambiguous particularly if one holds non-additive beliefs about outcome magnitude. The authors further pointed out that the state of being uncertain about whether a cue is causal or not would be expressed as intermediate likelihood ratings. Therefore, under the food allergist task, the tendency to give X middling ratings with low confidence makes X a moderate cause of the outcome, while the tendency to give Y low causal ratings with elevated confidence makes Y a decisive non-cause.

The argument about the role of causal ambiguity may be taken a step further by characterising the nature of uncertainty embedded in the redundant cues. Tannenbaum et al. (2017) classified uncertainty in judgement and decision making into two broad categories. Namely, uncertainty about the causal status of a cue is said to be epistemic if the probability of an outcome following is potentially knowable but concealed from the learner, and is said to be aleatory if the probability of an outcome following is random, manifesting as stochastic variability around a known base-rate. This distinction is particularly relevant to the implementation of different reinforcement schedules in the redundancy effect design. In the standard preparation, X is consistently paired with the outcome but its causal status may be rendered ambiguous, concealed from the learner by the presence of A. In the case of Y, even though Y is intermittently paired with the outcome, accurate predictions for the two trial types are achievable by referring to the reliable predictors that co-occur with Y. However,

causal ambiguity for Y would similarly be high if the accompanying B is deemed capable of preventing its outcome from occurring. For both cues, causal ambiguity does not arise from the inherent unpredictability of the outcome but bears upon the plausibility of alternative explanations under *a priori* assumptions. Where only one possibility is consistent with existing beliefs, there is no causal ambiguity at all. For example, emphasising additive assumptions for X and non-preventative assumptions for Y should readily allow learners to disambiguate the redundant cues as both being non-causal. Where more than one possibility is consistent with prior assumptions, the causal status of the redundant cues would be ambiguous. For example, two equally likely possibilities exist when learners entertain non-additive assumptions for X and preventative assumptions for Y, that is, the redundant cues can either be causal or non-causal. The type of causal uncertainty for X and Y under the standard redundancy effect training is thus of an epistemic nature.

The nature of uncertainty does not remain the same when the reinforcement schedules for the redundant cues change. Uengoer et al. (2013) invoked a partial reinforcement schedule for the blocking procedure (A+/–, AX+/–), pairing X with both the presence and absence of the outcome just as Y was in a relative validity (BY–, CY+), and found greater learning about X over Y regardless. This design balances X and Y with respect to their outcome probability, but does not equate X and Y with respect to their outcome predictability. Although the outcome rate is 50% for both cues, the outcome occurs randomly 50% of the time for X but can, in principle, be accurately expected on trials involving Y. In other words, the outcome is not only unstable but unpredictable for X and is equally unstable but predictable for Y. Besides epistemic uncertainty which may be present for both cues based on learners' prior beliefs, the impossibility to solve AX+/– trials creates an additional source of aleatory uncertainty for X.[3] Any difference in learning between X and Y observed with this and other designs using a probabilistically trained X would thus be open to

[3]Note that this is also true of the typical pseudo-discrimination control for the uncorrelated cue.

interpretation in terms of the different type of uncertainty highlighted by the training schedule.

Protecting Beliefs about Known Causes and Non-causes

As discussed above, there are reasons why causal judgment of redundant cues may be inversely related to the certainty with which the causal belief is held, and this pattern has been reported as least once in relation to the redundancy effect (Jones et al., 2019). In this context, the *dissociative pattern* observed involves the blocked cue being judged as a more likely cause but with lower confidence, and the uncorrelated cue being judged as a less likely cause but with higher confidence. Whether this dissociation is achieved through premises inherently embedded in the task scenario or assumptions explicitly manipulated in the experimental design, it informs an interesting question about how causal beliefs shape future learning. Recent research has investigated the subsequent fate of these redundant cues in the face of an incompatible new outcome: will future learning be guided by the difference in causality judgment or prediction certainty? While the former implies learning driven by prediction error, the latter implies learning driven by causal ambiguity.

The *theory protection* hypothesis makes divergent predictions from that of a prediction error account. Spicer et al. (2020, 2022) articulated the idea that, when faced with new information that conflicts with existing causal knowledge, humans are inclined to protect well-established beliefs from potential violations by attributing the incompatible information to the occurrence of other co-present cues with a less certain causal status. Chow et al. (2024) investigated the protective influence from an unknown cue during the extinction of an established cause. The unknown cue was either a hidden cue, which was physically absent but could be inferred from instructions, or a novel cue, which has never been shown in previous training. The authors found that the extent to which causal beliefs about the target cause were protected from extinction was related to the degree to which the unobserved cue

became preventative. This result suggests that theory protection about a known cause depends on the plausibility of the alternative cause as a preventer.

Now, consider a situation where X and Y, having been trained as blocked and uncorrelated cues, respectively, are then simultaneously paired with outcome presence. The occurrence of the outcome following the XY compound is not expected on the basis of the implausible cause Y, and is not fully predicted by the ambiguous X either. According to the theory protection principle, a greater update in causal knowledge is expected to take place for the cue associated with greater uncertainty. In this case, the more causally ambiguous X should undergo greater update than the unambiguously non-causal Y. Consistent with theory protection, Spicer et al. (2020) found more updating of X than Y using this compound training in the food allergist task.

The asymmetrical learning about components of the same compound is difficult to reconcile with error-correction mechanisms central to many traditional associative models. Theories that rely on an individual error term predict greater learning about the cue that engenders a larger prediction error. In the case of the XY compound, Y should undergo a larger associative increment than X because Y has weaker associative strength than X (at least according to the causal ratings that participants typically provide), and thus Y carries a larger individual prediction error than X, when the outcome is experienced. This is the opposite of what was empirically found by Spicer et al. (2020). Theories based on common error learning rules anticipate equivalent learning for all co-present cues and thus also struggle to explain this finding. Rescorla (2001) conceived of the possibility of elaborating the common error term with appropriate function mapping associative change onto performance. Holmes and colleagues (2019) formalised this idea by postulating a non-linear double sigmoid function that translates changes in associative strength to performance depending on the initial associative values of the constituent cues of a compound. This

function successfully captures the X learning bias observed by Spicer et al. (2020), providing a purely associative explanation for the fate of redundant cues in future learning. However, Spicer et al.'s study represents the only instance to date in which the theory protection idea was interrogated using the redundancy effect, and an allergist task was used in this study. The strong possibility remains that different results would be obtained with tasks that do not invite strong non-preventative assumptions.

Chapter Summary

The redundancy effect represents a significant challenge to the traditional associationist approach as none of the mechanistic theories proposed thus far can provide a satisfactory account of the empirical findings. Chapter 2 will glean insight into this seeming conundrum by dissecting promising associative models into two components. The first is the fundamental error-correction algorithm that governs associative change according to either a separable or a summed discrepancy between the prediction based on the cues present and the experienced outcome. The second is the rule for attention deployment that provides an additional control over the extent to which cues are processed based on their predictive utility relative to co-occurring competitors. Through manipulations of outcome probability and predictability, the validity of three theoretical constructs, absolute cue-outcome relationship, relative informativeness, and outcome predictability, are evaluated in relation to formal associative theories. The findings of Chapter 2 provide critical tests of the capability of current associative learning models to explain learning about ambiguous cues.

Human causal learning appears to involve controlled and effortful deduction. This type of logical inferencing is clearly a complex cognitive function that is not captured by elementary associative models. It is, however, worth exploring whether inferential reasoning as informed by learners' preconceived assumptions is implicated in the judgement of redundant cues. If so, it may offer a better explanation of the redundancy effect and related

phenomena, especially in circumstances where associative models do not fare well. Causal ambiguity around the blocked and uncorrelated cues has been shown to be related to the ease with which deduction of non-causality can be applied based on *a priori* assumptions. In particular, magnitude additivity assumptions validate the inference that the blocked cue cannot be causal. Likewise, non-preventative assumptions allow for deductive reasoning about the uncorrelated cue being non-causal. By manipulating prior assumptions in a direction that either encourages or discourages deductive inference about redundant cues in a pretraining phase, Chapter 3 will shed light on the propositional processes involved in human causal reasoning about ambiguous cues, which vary with individual propensity to engage in different learning processes.

The ability to form strong causal beliefs following logical deduction enables further inquiry into the fate of the blocked and uncorrelated cues in future learning. Given that learners may rely on complex reasoning processes to disambiguate the causal role of redundant cues, it is conceivable that such reliance on deduced inference may persist as a cognitive heuristic to influence causality judgment in future learning. However, under conditions where deductive reasoning can be more readily applied for one of the redundant cues than the other, the two critical cues are predicted to not only differ in causal judgment but also in confidence in the reversed direction. While both being consequences of deductive reasoning, it remains to be seen whether causal judgment difference or confidence difference would exert a stronger influence on subsequent learning, especially when X and Y are simultaneously presented as a compound. In line with the theory protection proposal (Spicer et al., 2020), new learning is expected to be biased toward the causally ambiguous X in order to protect existing knowledge about Y being non-causal. The experiments in Chapter 4 implement a similar pretraining phase to manipulate learners' prior beliefs regarding magnitude additivity and preventative relationships. Following the compound test procedure,

the differential readiness to engage in inferential reasoning will be shown to modulate the tendency to maintain established beliefs about redundant cues in future learning.

The inherent nature of cover story may bias deductive inference in ways that are beyond the influence of experimental manipulations. Across all current experiments, the relatively new hormone change task was used. This task has the advantage of making preventative cue-outcome relationships plausible, which are particularly relevant for causal judgment about the uncorrelated cue. However, a considerable body of past research on the redundancy effect (e.g. Uengoer et al., 2013; Uengoer et al., 2020) has by convention adopted the food allergist task where prevention is seen as virtually impossible based on real-life interactions with allergenic foods. The potential inconsistencies that might arise from using different task scenarios behoove us to compare the two task paradigms directly in a single study. The experiments in Chapter 5 present contingency information for the redundancy effect and the relative validity effect (i.e. effects that critically depend on learning about the uncorrelated cue) in summary format under both the food allergist scenario and the hormone change scenario. The findings will allow better understanding of the current empirical results in relation to studies that used a different cover story.

Chapter 2

Learning about Redundant Cues with Uncertain Outcomes

The day-to-day environment contains a diverse array of causally connected events. The complexity of these relationships necessitates decision making based on ambiguous evidence. Our ability to infer causal relationships in uncertain situations has enabled humans to predict ensuing consequences, make sensible judgments, and adequately adjust behaviours to accomplish goals and avoid punishments. Understanding how and why ambiguous events influence our judgments of causality in certain circumstances but fail to do so in others is thus important. In recent years, inquiries into the fate of different redundant cues that act as ambiguous signals of important outcomes have been brought into focus by a phenomenon known as the redundancy effect (e.g. Pearce et al., 2012; Uengoer et al., 2013).

The redundancy effect describes the finding of stronger outcome anticipation elicited by the blocked cue X than that elicited by the uncorrelated cue Y. Uengoer et al. (2013) attempted to clarify the mechanism through which this phenomenon is achieved by contrasting two differences inherent in the training history. First, the authors argued that while both X and Y are considered redundant in terms of their usefulness in predicting the outcome, they differ in the informational value that they possess relative to the cues accompanying them. In a blocking procedure, although A is more frequently associated with the outcome than X, both A and X always lead to the outcome on their presentations. The usefulness of X is thus discounted by the fact that it conveys no additional, but rather the exact same, information as A. In contrast, the informativeness of the cues varies among cues Y, B and C (i.e. those involved in a *relative validity*). The intermittent pairing of Y with the outcome places it at a disadvantage compared to B and C; Y serves as a poorer signal of the absence of the outcome than the continuously non-reinforced B and as a poorer signal of the presence of the outcome than the continuously reinforced C. When X and Y are presented

individually at test, X, which has been equally informative as its associate A, is regarded as a more likely cause of the outcome than Y, which conveyed less information than its associates B and C. Following from this comparison of relative informativeness, the redundancy effect may be explained as the result of Y being less relevant relative to the accompanying B and C than X is relative to A at the time of learning.[4]

Table 2.1

Key design elements and findings in Uengoer et al. (2013)

Training	Expected finding
Blocking: A+ AX+ DE+	Blocking effect: X < E
Relative Validity: BY– CY+ FG+/– HG+/–	Relative Validity effect: Y < G
	Redundancy effect: X > Y
Blocking (partial rft): I+/– IZ+/–	Redundancy effect (partial rft): Z > Y

Uengoer and colleagues (2013) further argued that X and Y not only possess different relative utility, they also differ in terms of the absolute relationship they bear with the outcome. This is especially the case where the absence of the outcome is assumed to be mentally represented as a distinct event in and of itself. That is, X is always followed by the "outcome present" event (in Uengoer et al.'s case an allergic reaction), whereas Y is sometimes followed by an "outcome absent" event (e.g. no allergic reaction). When cues are tested after training, Y elicits a weaker outcome expectation because it has been experienced as a non-predictor of the outcome compared to X which has never been presented without the

[4]In addition to an important role for relative informativeness during training, memory retrieval of the paired cues via within-compound associations may act as a supplementary process that exacerbates the difference between X and Y further at the time of testing. That is, X which brings to mind an equally valid predictor A may elicit a greater expectation of the outcome than Y which activates the representation of the more informative companions B and C.

outcome. As Uengoer et al. (2020) later pointed out, the different learned contingencies can result in greater associative strength being accrued to X than to Y if incorporating an individual error learning algorithm. The absolute relationship explanation therefore attributes the difference in likelihood judgments for X and Y to the difference in their reinforcement schedule irrespective of the cues with which X and Y are paired.

In essence, two properties of the statistical contingencies in the learning task of X and Y were envisaged in Uengoer et al. (2013) as critical factors for the redundancy effect. These factors are both in line with the higher outcome expectation associated with X than with Y. On one hand, the redundancy effect can be understood by taking into consideration the relative informativeness of cues where comparison of informational value provided by the target redundant cue and its associates determines the likelihood judgment. Alternatively, a more straightforward account is possible through inspection of the reinforcement schedule under which the cues are trained. It is instead the difference in experienced contingencies independent of the influence of other cues that leads to differences in encoded and retrieved memories about training trials.

Uengoer and colleagues (2013, Experiment 3) investigated the absolute relationship proposal by equating the outcome rate for X and Y at 50%. Specifically, A and AX trials from the blocking treatment were followed by the outcome probabilistically at a 50% rate to match the outcome presentation on BY and CY trials from relative validity, rather than being followed by the outcome deterministically (at 100%) as in an archetypal design. This design ensured that both X and Y have been involved in outcome-present and outcome-absent trials and were therefore capable of activating the representation or retrieving memories of the outcome and its non-occurrence. At the same time, X remains equally predictive as A, whereas Y remains less predictive than B and C. If the difference in absolute relationship that X and Y hold with the outcome in the typical design was the key reason why X elicits

stronger predictions than Y, then this difference should be eliminated when the probability of the outcome given cue X, i.e. p(O|X), and p(O|Y) both equal 0.5. However, X continued to be regarded as a more causal cue than Y despite equivalent absolute relationships. The authors concluded that the different absolute relationships with the outcome for cues X and Y cannot be responsible for the redundancy effect, yet the relative informativeness of the target cues remains a likely factor for the phenomenon. Table 2.1 outlines the key contingencies of note from Uengoer et al.'s experiment. Note that the partially reinforced blocking design is represented using cues I and Z (i.e. I+/– IZ+/–) to distinguish it from the typical continuously reinforced blocking design (A+ AX+) since both appear in the same within-subjects design in several experiments.

Alongside the observation of the redundancy effect with a partially reinforced blocked cue, Uengoer et al. (2013) documented an unusual pattern of causal judgment in relation to the blocking effect under a probabilistic setting. When trained at 50%, the authors found that the blocking A and the blocked X were regarded as being equally causal of the outcome. This raises the question of whether a lack of competition between the concomitant cues itself could provide a simpler explanation for the high causal judgment for X in this case. However, given the rather scarce empirical data on the blocking effect with an inconsistent training schedule, it is difficult to establish the reliability of this single observation. One of the aims of the current study was therefore to unveil more clearly the role that absolute relationship with the outcome plays in the redundancy effect.

The blocking effect has been a benchmark for the development of associative learning theories. Many of these theories that predict restricted learning about the blocked cue in its prototypical form also generate similar predictions about the blocked cue trained with a probabilistic outcome. For example, the influential Rescorla-Wagner model (Rescorla & Wagner, 1972) and modified versions that have been applied to account for the redundancy

effect (Vogel & Wagner, 2017) both predict that X will acquire relatively little associative strength in the presence of an equally reliable alternative predictor, A, that is also trained by itself. They assert that the amount of associative change on a given trial is governed by the discrepancy between the total outcome expectation elicited by all cues present and the observed outcome (i.e. summed prediction error). So long as the asymptotic associative strength supportable by the outcome remains unchanged with the addition of X to A, the negligible prediction error produced on AX trials should result in very small, if any, increment in learning about X. The lack of judgment difference between A and X with a 50% pairing schedule thus appears to be at variance with these models that account for blocking. However, the difficulty may be partly resolved by entertaining an additional assumption about the degree of *commonness* across training cues. For example, the context of performing a learning trial (i.e. the perceptual elements that are common to all trials) may become strongly associated with the outcome, preventing associations being formed from both A and X to the outcome when trained at 50%. Whereas animal studies have a tradition of splitting trial commonness over both the shared stimulus elements and the context (i.e. especially the silent period within the inter-trial interval), we combined both as the general common context of the experiment that was kept constant across trials for computational simplicity.

As Uengoer and colleagues (2019) have showed, predictions derived from other models that ascribe cue interaction to the relative informativeness of cues via changes in attention may fare well with the foregoing results. Taking the Mackintosh model (Mackintosh, 1975) as an example, it is capable of predicting the critical judgment difference between X and Y, as well as blocking and other competitive learning effects by virtue of a selective attention mechanism that focuses attention on valid predictors while filtering out poor ones. Importantly, this theory anticipates that A and X will remain similarly associable under the 50% intermittent pairing schedule with the consequence of both cues possessing

similar associative strength at any point during training, thus resolving the puzzle around equal judgment for A and X under partial reinforcement. Building on from Mackintosh (1975), Uengoer and colleagues (2020) in a recent study devised a new associative learning model which integrates separable error learning algorithms with an attention change mechanism akin to that of the Mackintosh theory. The separable error term not only accounts for the lack of blocking with a 50% contingency, but also makes successful predictions about the redundancy effect itself on the basis of greater cue-outcome pairings for X than for Y. The associability of A and X is predicted to alternate between increasing and decreasing under partial reinforcement, which is then combined with the individual error term to jointly result in comparable levels of learning for both cues.

It is note-worthy that the choice of cover story may have a bearing on likelihood judgment for Y. The conventional food allergist task requires participants to predict allergic reactions from certain combinations of foods eaten. Several authors have revealed a down-modulating effect of this paradigm on likelihood ratings to Y (Jones et al., 2019; Zaksaite & Jones, 2020). It is uncommon in real world situations that foods can prevent an allergic reaction from occurring. As a result, learners appear to disregard any potential preventative influence from foods consumed together with Y (i.e. B) and take the absence of the outcome on BY trials as strong evidence of both B and Y being non-allergenic. This line of reasoning would not affect judgments of X because X is always followed by the outcome. Even in conditions where AX is only probabilistically followed by the outcome, it is unlikely a participant would discount X in the same way as Y because the participant never has the opportunity to observe the effect of X without A (in contrast, the participant has the opportunity to observe that Y has no effect in the absence of C). Given that the redundancy effect involves learning about the absence of an outcome where one might expect it, the effect might be particularly sensitive to beliefs about preventative relationships. As a

consequence, there is strong need to replicate any key finding shown with the food allergist task, using other learning tasks that do not convey the same non-preventative prior beliefs. Uengoer's result using partial reinforcement in blocking is a clear example of this; the particularly low likelihood judgment for Y might have resulted from the food allergist task itself, thus generating a difference between X and Y for reasons that are beyond the scope of associative learning models.

From the modelling analyses conducted in previous studies, successful prediction of an enduring redundancy effect under probabilistic conditions appears to be what distinguishes the credibility of these associative models. However, the X > Y result has only been demonstrated under partial reinforcement on a single occasion. Given the importance of this result for determining the capability of learning theories, more evidence is needed to understand how redundant cues trained in procedures that give rise to differential cue predictiveness and outcome predictability are treated in causal learning.

In the current study, we provide a more extensive investigation of changes in the redundancy effect when outcome uncertainty is introduced on blocking trials. We report a series of four experiments, followed by simulations with variants of error correction learning algorithms that purport to capture the redundancy effect in different ways. The experiments used a learning task where preventative and generative relationships are both plausible; a variant of the hormone change task where participants played the role of a medical researcher, attempting to predict hormone changes in patients based on medicines taken (Zaksaite & Jones, 2020). Experiment 2.1 used Uengoer et al.'s original partial reinforcement design in an attempt to replicate the redundancy effect with partially reinforced blocking contingencies. Experiments 2.2 and 2.3 then examined the impact of varying the base-rate probability of the outcome in different ways, by controlling the base rate at 50% with the addition of filler trials (Experiment 2.2), or by presenting a 'no cue' trial on which the

outcome was present at the rate of either 0% or 50% (Experiment 2.3). Having found a surprising and yet fairly consistent effect of partial reinforcement across these three experiments, in which the redundancy effect was *not* observed when the blocking contingencies were partially reinforced, in Experiment 2.4 we examined whether the introduction of even a relatively small amount of outcome uncertainty into the predictive contingencies prevents the redundancy effect. The results of these experiments, which paint a very different picture of the necessary conditions for observing the redundancy effect, were then simulated using associative learning algorithms that use either a summed or separable error term, and assume attention to individual cues is either fixed or variable (via either the Mackintosh or Uengoer model algorithms).

Experiment 2.1

The first experiment sought to provide a replication of Experiment 3 of Uengoer et al. (2013) which included trials of the kind summarised in Table 2.1. This design included a probabilistically trained blocking contingencies (I+/–, IZ+/–) which were compared with a set of prototypical blocking contingencies (A+, AX+) in terms of their ability to elicit the redundancy effect. Experiment 2.1 adopted the same within-subjects design as Uengoer et al. but within a different cover story (i.e. the hormone change task). The authors of the original paper found evidence of greater outcome expectation for Z than for Y after equating the outcome probability at 50% for both redundant cues. This finding seems to eliminate an explanation for the redundancy effect in terms of the relationship that X and Y hold with the outcome, and favors the difference in the informational value that X and Y each possesses relative to their associates as the main driving factor. However, the similar likelihood judgments for blocking and blocked cues under partial reinforcement suggests that ratings for Z may not have been affected by cue competition at all, which would place the results at odds with most of the associative learning explanations described so far. This calls into question

the generality of the finding. Before further exploration of the theoretical significance of this result can be carried out, it is necessary to confirm that the Z > Y effect is generalisable beyond the specific experimental parameters used by Uengoer and colleagues in their single demonstration.

Method

Participants

The participants for this online study constituted both first-year psychology students from the University of Sydney and Prolific Academic participants. The former cohort received course credits and the latter cohort received monetary rewards in return for their participation. A total of 61 participants were recruited with 11 excluded for failing to pass the instruction check question, the write check question, or the training criteria. The remaining 50 participants (25 females, mean age=24.25, *SD*=7.89) were included in the subsequent analyses. The sample size was estimated based on Uengoer et al. (2013, Experiment 3) where 20 participants were recruited for a single group experiment. The current sample achieved over 90% power to detect any meaningful difference in learning between the blocked and uncorrelated cues of effect size dz=0.5.

Design

The design of this experiment is summarised in Table 2.2. Participants began with a training phase during which causal relationships between various medicines and hormone changes were learned. Causal knowledge was then assessed in the likelihood ratings tests, as well as an additional forced choice test administered after training.

Table 2.2

Design of Experiment 2.1

Training	Likelihood ratings test	Forced choice test
A+ AX+ BY– CY+ I+/– IZ+/–	X Y A B C I Z	X vs. Y Z vs. Y X vs. Z

Note. Letters represent cues, randomly assigned to different medicine names. "+" represents hormone increase and "–" represents no hormone change.

Apparatus and Stimuli

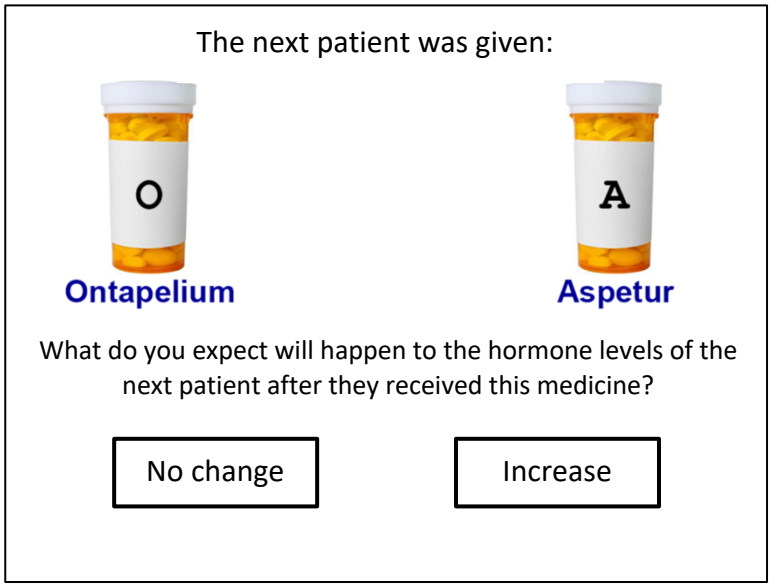
The experiment was programmed using jsPsych and completed on desktop computers or laptops. The experimental cues were six 300×300 pixel images of medicine bottles. Below each bottle was a fictitious medicine name written in blue. These names included *Aspetur, Broncin, Chrurin, Dioxnyl, Ephemerol, Felicium, Gambutrol, Hyronalin, Impbatine, Jamitol, Krayoxx, Lithorol, Metazine, Nozambutol, Ontapelium, Plycidox, Quelinum, Rodvuccial and Sarizol*. To help distinguish among medicine names, medicine bottles were labelled with the first letter of the medicine names in black. Medicines were randomly assigned to cues in Table 2.2. Cues were assigned a drug name, by randomly sampling from the pool of drug names without replacement. The binary outcomes were 'no change' and 'increase' in hormone levels presented as black text inside rectangular boxes. At the start of each training trial, one or two cues appeared on the upper half of the screen and the outcomes appeared on the lower half of the screen. After each prediction, corrective feedback was provided at the centre of the screen as 'correct' written in green if the right decision was made or 'incorrect' written in red if the choice was wrong, with the correct outcome appearing below. The entire experiment was presented against a white background. Participants read through instructions by pressing the space bar and responded by clicking on the chosen outcome.

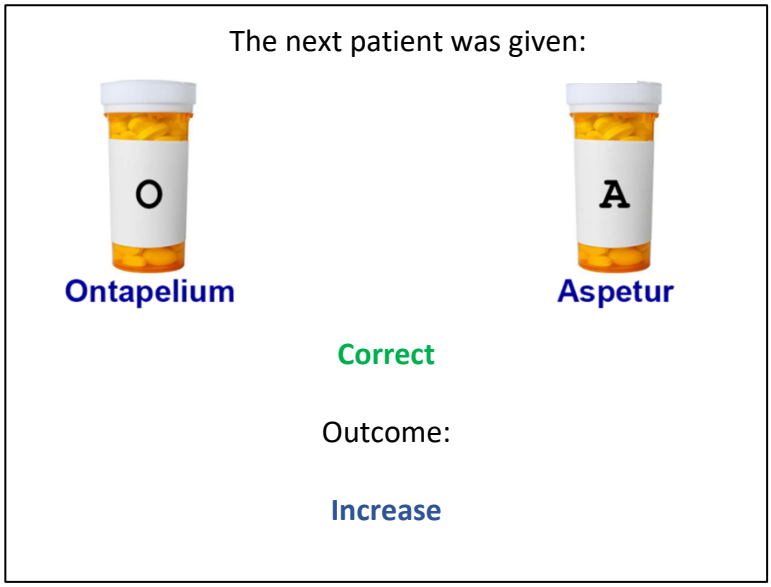
Procedure

Figure 2.1

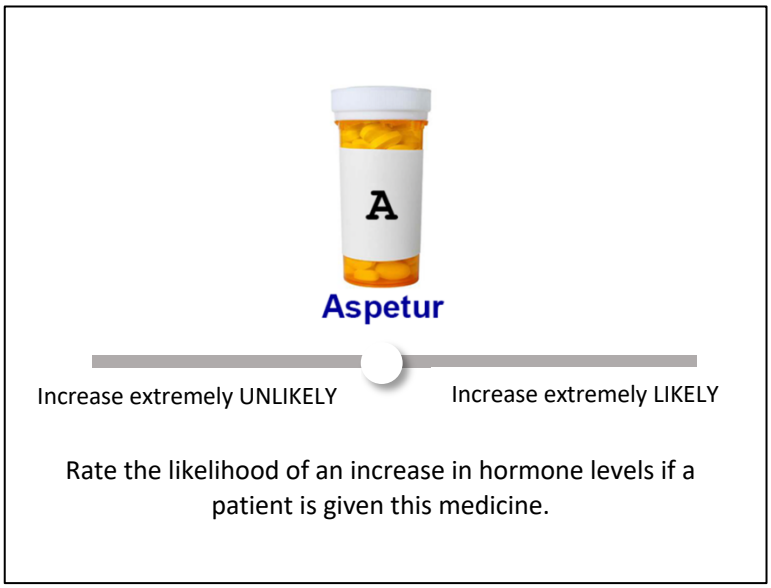
Schematic diagrams of (a) the training phase, (b) the likelihood ratings test, and (c) the forced choice test.

(a)

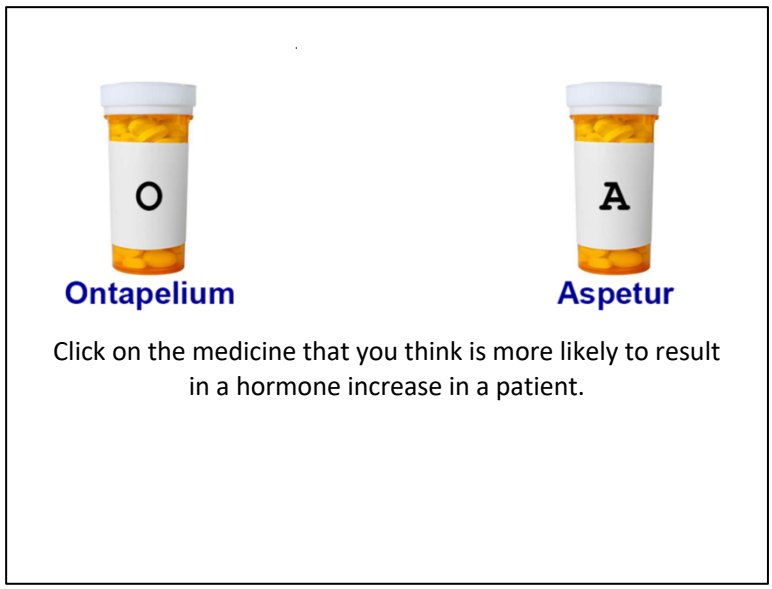




(b)



(c)



The procedure of the experiment is depicted in Figure 2.1. At the beginning of the task, participants were asked to imagine themselves to be a medical researcher, who is interested in studying the effect of different medicines on hormone change. Their task was to predict whether medicines that Patient X consumed would cause no change or an increase in their hormone levels. The training phase comprised 10 blocks of trials. Within each block, each of the 6 trial types (A+, AX+, BY-, CY+, I+/-, and IZ+/-) were presented twice in a randomised order. On each training trial, participants were presented with either one or two cues and decided whether an increase in hormone levels would result after taking the medicine(s) by clicking on the corresponding outcome option. The spatial location of the cues was counterbalanced for two-cue trials such that each cue was displayed once on the left side and once on the right side of the screen.

Participants moved onto the test phase immediately after training. Two types of tests, the ratings test and the forced choice test, were administered to provide a measure of how likely participants believed that each medicine would cause a hormone increase. On the ratings test, medicines were shown individually one after the other and participants were instructed to rate the likelihood that the given medicine would lead to an increase in hormone levels based on their knowledge acquired during training. Participants indicated their responses using a sliding scale from ‘Increase extremely UNLIKELY’ to ‘Increase extremely LIKELY’. The cues were presented on the upper half of the screen and the scale was presented underneath the cues. The forced choice test was given after the ratings test, where cues were compared against each other in pairs and participants selected one cue from each pair that they believed would be more likely to result in a hormone increase. Cue pairs each appeared four times and cue positions were counterbalanced for left and right.

Data Analysis

Training accuracy was determined on the basis of trials which had a decisively accurate outcome. That is, trials with a probabilistic outcome were excluded as neither the presence nor the absence of the outcome can be seen as accurate on these partially reinforced trials. Learning scores were calculated by subtracting the number of incorrect outcome choices from the number of correct outcome choices. A positive learning score resulted if the number of correct responses was greater than the number of incorrect responses, and a greater number of incorrect choices produced a negative score. Participants who failed to achieve a positive learning score in the last half of training, failed the instruction check more than three times, or admitted having written notes during the experiment were excluded from analyses. The likelihood ratings scale was calibrated from 0 to +100 with more positive values indicating higher likelihood of causing the outcome and less positive values indicating lower likelihood of causing the outcome. The binary nature of the response on the forced choice test implies potential violations to the Gaussian probability distribution of sample data assumed by common parametric statistical procedures. For this reason, the non-parametric sign test was employed in the analyses of cue choice data where probability for choosing the first cue from each pair is coded as 1 if greater than 50%, 0 if equal to 50%, and -1 if less than 50%. Sign tests were one-tailed and all other tests were two-tailed with significance set at 0.05.

Results

Training

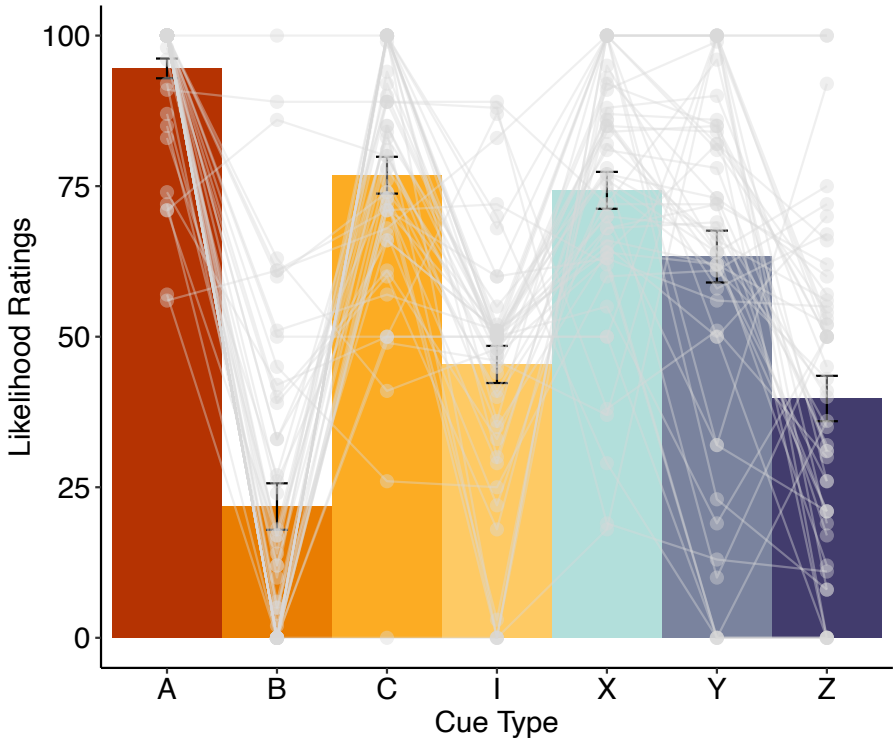
Learning of cue-outcome relationships proceeded fairly rapidly for trials with a definitive outcome (please see supplementary materials for training graphs). Participants reached an average accuracy of 91.94% in the last half of training. A (4x10) repeated measures ANOVA with trial type and block (1-10) as within-subjects factors revealed a

significant quadratic trend, $F(1,48)=46.13$, $p<.001$, $\eta_p^2=.490$, confirming the rapid increase in prediction accuracy across training blocks.

Likelihood Ratings

Figure 2.2

Mean likelihood ratings on the ratings test of Experiment 2.1. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.



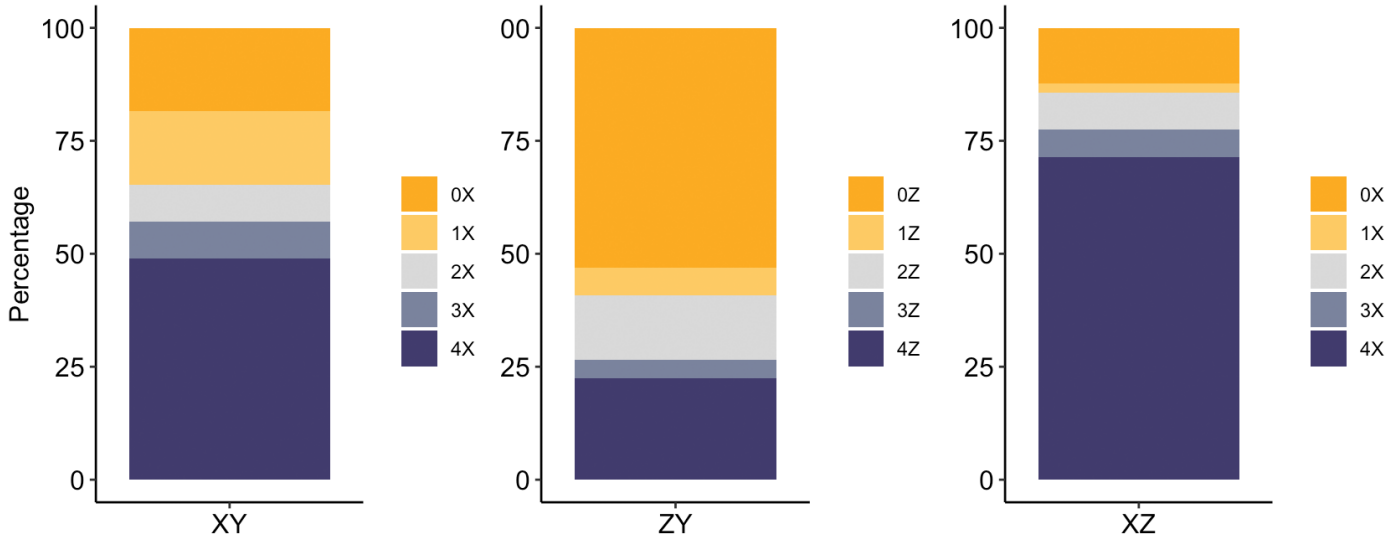
Performance on the likelihood ratings test as illustrated in Figure 2.2 was analysed through paired samples t-tests. The four comparisons of main interest were between X and Y, Z and Y, A and X, as well as between I and Z. The blocked cue X was rated as a significantly more likely cause of the outcome than the uncorrelated cue Y, $t(48)=2.52$, $p=.015$, $d=.359$. However, this characteristic pattern of the redundancy effect was reversed when the blocked cue Z was followed by the outcome on only 50% of its presentations, $t(48)=4.86$, $p<.001$, $d=.695$. The blocked cue X was regarded as a significantly less likely cause than the blocking cue A, $t(48)=6.40$, $p<.001$, $d=.914$. This difference was non-significant between I and Z,

t(48)=1.05, *p*=.301, *d*=.149, indicating a lack of effective blocking under a partial reinforcement schedule.

Forced Choice

Figure 2.3

Mean percentage of choosing the first cue from the XY pair, the ZY pair and the XZ pair on the forced choice test in Experiment 2.1. Higher percentage indicates higher likelihood that a given cue is chosen as a cause.



Percentage of choices from four repeated presentations of each cue pair on the forced choice test is depicted in Figure 2.3. Sign test results indicated that although X had a slightly higher probability of being chosen as the more likely cause than Y, this difference was not statistically significant, *p*=.067. Z was less often chosen than Y, *p*=.009, supplementing the reversed redundancy effect revealed through likelihood ratings. X was chosen significantly more frequently than Z, *p*<.001, which also supports the result from the ratings test.

Discussion

The purpose of Experiment 2.1 was to provide a direct replication of the important results reported by Uengoer and colleagues (2013), using the exact same design but in a

different causal judgment task. Consistent with the standard redundancy effect, the consistently trained blocked cue X was judged as being a more likely cause of hormone increase than the uncorrelated cue Y. In contrast, the 50% partially reinforced blocked cue Z was rated as a less likely cause than Y. The supplementary forced choice test offers additional support for the reversal of the redundancy effect between Z and Y, but provides less clear evidence for the learning difference between X and Y. As described in previous sections, Z can be seen as equivalent to Y in terms of outcome probability, but different to Y in respect of informativeness relative to concomitant cues. Assuming that learning is an increasing function of relative informativeness, the reversed effect is the opposite of that expected by the higher relative informational value of Z over that of Y.

This result is clearly at odds with Uengoer et al.'s findings when the equivalent design was presented under the guise of the widely used food allergist task. However, we will save discussion of the impact of the learning scenario to the General Discussion, as Experiments 2.2 and 2.3 were motivated by another aspect of the design of Experiment 2.1. The current (Z < Y) result appears to reflect the low outcome probability of the blocking procedure under partial reinforcement compared to the overall outcome rate. The context may be having role in this case. The high base rate of 67% may establish the context as a strong predictor to compete for associative strength with the 50% partially reinforced I and Z on reinforced trials, thereby retarding learning about both cues which indeed signal a 17% drop in outcome probability. If we assume that participants associate the I/Z cues with a reduction in the rate of the outcome, the blocking treatment should presumably restrict learning about Z as being a conditioned inhibitor by establishing I as a better signal of this reduction in outcome probability. However, consistent with the lack of blocking observed in the original paper, the inhibitory influence of Z on the outcome was not hindered by the supposedly stronger inhibitor I (i.e. elevating associative strength for Z), resulting in low likelihood judgments for

both cues. The uncorrelated cue may be less affected in this sense because the context better signals outcome occurrence on CY trials but predicts outcome non-occurrence on BY trials worse.

The possible impact of the base-rate on learning about ambiguous cues motivated the changes introduced in Experiments 2.2 and 2.3, as we sought further evidence for the redundancy effect under partial reinforcement conditions.

Experiment 2.2

One atypical feature of the design used by Uengoer et al. (2013; Experiment 3) and in our Experiment 2.1 is that under partial reinforcement, the blocking cue signals a *reduction* in the probability of the outcome relative to the base-rate; the chance of outcome occurring following the 50% partially reinforced blocked cue was lower than the 67% overall outcome rate in the task. While Uengoer's orginal result (Z > Y) runs opposite to the significant difference that we observed (i.e. Z < Y), it is possible that at least one of these results occurred because of the influence of this design feature. For instance, the hidden effect of this unusual training could have been made more obvious by the hormone change task which allows Y to retain a weak causal relationship with the outcome in a way that the food allergist task may not. Experiment 2.2 therefore compared the same critical ambiguous cues used in Experiment 2.1 but embedded within a more complex design in which the overall outcome rate was 0.5.

Method

Participants

Eighty-four undergraduate psychology students from the University of Sydney participated in this experiment for partial course credit. Twelve were removed on the basis of the exclusion criteria, leaving seventy-two in the final analyses (56 Females, mean age=20.21, *SD*=2.01).

Design and Procedure

The design of Experiment 2.2, shown in Table 2.3, included the overshadowing controls for blocking under consistent (DE+) and intermittent (JK+/–) reinforcement schedules, as well as filler trials (L–, MN–, and OP–) to control the overall outcome rate at 50%. The overshadowing controls were compared against the blocked cues in the likelihood ratings test and the forced choice test. The procedure was the same as per Experiment 2.1.

Table 2.3

Design of Experiment 2.2

Training	Likelihood ratings test	Forced choice test
A+ AX+ BY– CY+ I+/–	X Y Z A B C D E F G	X vs. Y X vs. Z Y vs. Z
IZ+/– DE+ FG+/– GH+/–	H I J K L M N O P	X vs. E Y vs.G Z vs. K
JK+/– L– MN– OP–		

Note. Letters represent cues, randomly assigned to different medicine names. "+" represents hormone increase and "–" represents no hormone change.

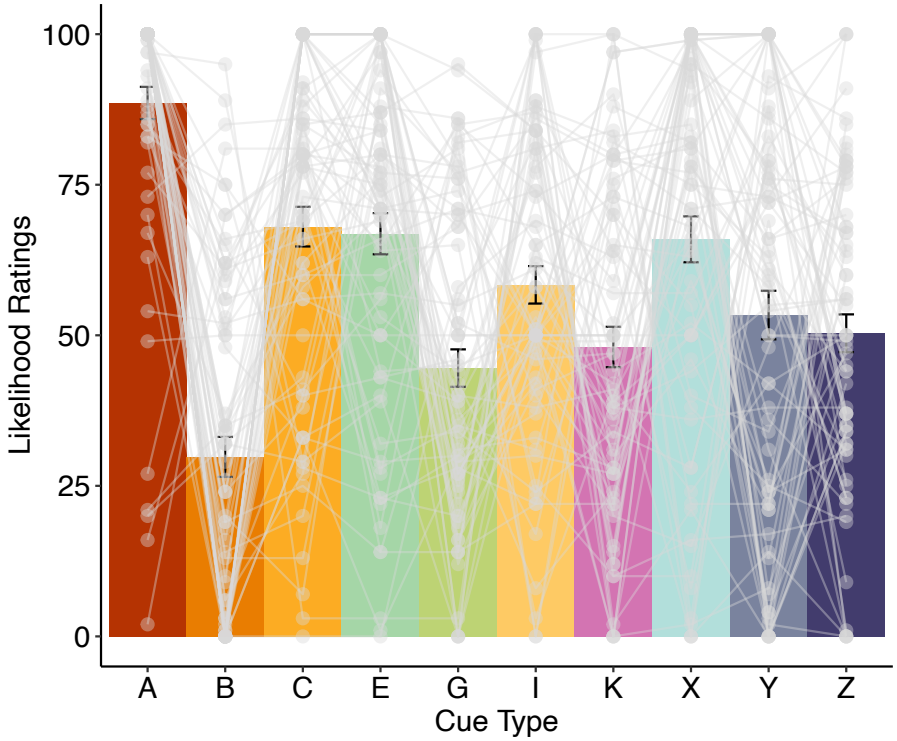
Results

There was a gradual improvement in performance across training for trials with a definitive outcome (please see supplementary materials for training graphs). Participants reached an average accuracy of 79.58% on trials with a predictable outcome in the last half of training. A (10x8) repeated measures ANOVA with block (1-10) and trial type as within-subjects factors revealed a significant quadratic trend for block, $F(1,71)=40.39$, $p<.001$, $\eta_p^2=.363$, indicating a rapid increase in training accuracy.

Likelihood Ratings

Figure 2.4

Mean likelihood ratings on the ratings test of Experiment 2.2. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.



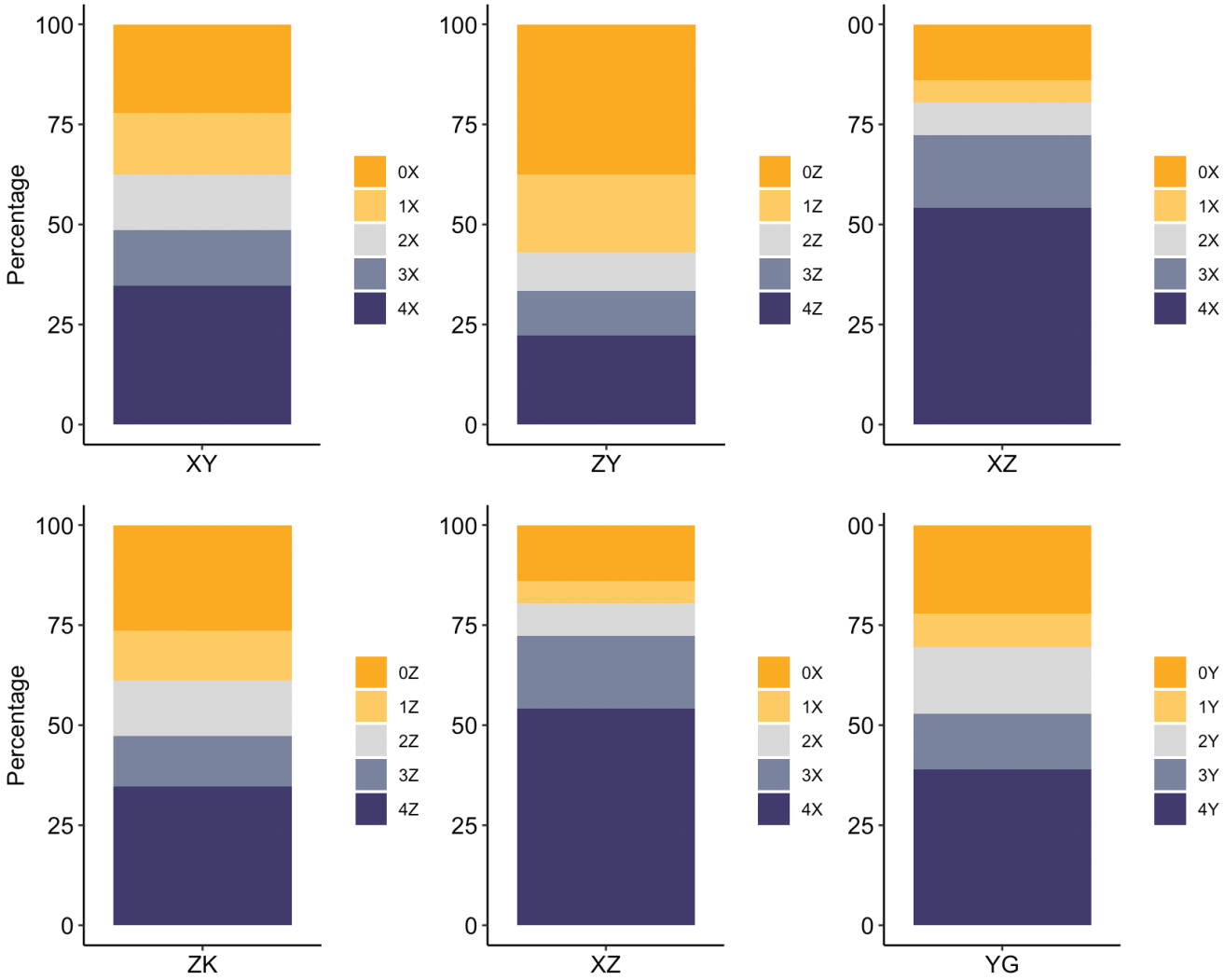
Participants' likelihood judgments were captured by the ratings test as shown in Figure 2.4. Please see supplementary materials for complete figures including all cues for Experiments 2.2-2.4. The main comparisons of interest were again between X and Y, Z and Y, and X and Z. However, comparisons between X and E, and between Z and K provide tests for the presence of blocking under continuous and partial reinforcement, respectively, and comparison between Y and G provides a test for the relative validity effect. Paired samples t-tests revealed significantly higher ratings for X than for Z, $t(71)=15.61$, $p=.001$, $d=.405$. Ratings for X were significantly higher than those for Y, $t(71)=2.37$, $p=.021$, $d=.279$, which is the hallmark finding of the redundancy effect. This pattern was however not observed between Z and Y when the blocked cue was intermittently reinforced at 50%, $t(71)=.61$, $p=.544$, $d=.072$. Contrary to the pattern expected, Y was rated numerically more likely than

G to cause the outcome though the difference did not reach significance, $t(71)=1.65$, $p=.104$, $d=.194$, indicating a lack of relative validity effect. Ratings for A were significantly higher than those for X, $t(71)=5.24$, $p<.001$, $d=.618$, but ratings for I did not differ significantly from those for Z, $t(71)=1.62$, $p=.110$, $d=.191$, consistent with the differences between the blocking cue and the blocked cue previously observed by Uengoer and colleagues (2013). Surprisingly, the blocking effect was neither significant for the X vs. E comparison, $t(71)=.20$, $p=.839$, $d=.024$, nor for the Z vs. K comparison, $t(71)=.49$, $p=.625$, $d=.058$.

Forced Choice

Figure 2.5

Mean percentage of choosing the first cue from the XY pair, the ZY pair, the XZ pair, the ZK pair, the XZ pair, and the YG pair on the forced choice test in Experiment 2.2. Higher percentage indicates higher likelihood that a given cue is chosen as a cause.



Percentage of choosing each target cue from four repetitions of each cue pair on the forced choice test is illustrated in Figure 2.5. Sign tests were conducted to compare within 6 cue pairs: XY, ZY, XE, ZK, XZ, and YG. Although the probability of choosing X was slightly higher than that of choosing Y, the redundancy effect was not significant, p=.187. The effect was reversed with Y being more often chosen than the intermittently reinforced blocked cue, Z, as the likely cause, p=.023. The blocking effect was however not significant either between X and E, p=.896, or between Z and K, p=.813. There was a higher chance of X being selected than Z, p<.001. Y was more frequently chosen than G, indicating a reversal of cue validity, p=.026.

Discussion

Experiment 2.2 found lower probability of the partially reinforced blocked cue Z being chosen as the more likely cause compared to Y on the forced choice test (i.e. the reverse of the redundancy effect) but a lack of judgment difference between the two kinds of redundant cues on the ratings test. These results were in keeping with the hypothesis that the learning difference between a probabilistically trained blocked cue and the uncorrelated cue hinges upon the underlying rate of outcome occurrence. Specifically, when the outcome probability following Z was lower than the base rate in Experiment 2.1, Z was regarded as being less likely to cause the outcome than Y as indexed by both likelihood judgment and forced choice. When the outcome probability and the overall outcome rate were made equivalent in Experiment 2.2, evidence of a reversed redundancy effect was less robust though it should be noted that the increased task complexity may have been partly responsible for this as well.

Experiment 2.3

The redundancy effect has previously been demonstrated in a condition where the overall outcome rate was slightly above the outcome probability for the 50% reinforced blocked cue within the food allergist task (Uengoer et al., 2013). The same training design led to a complete reversal of the effect with a shift in cover story to the hormone change scenario in Experiment 2.1. Further manipulation to control the base rate at chance resulted in an absence of the effect in Experiment 2.2. In Experiment 2.3 we attempted to manipulate outcome base rate in a complementary fashion, by introducing 'no cue' trials in which the base rate of the outcome without intervention could be observed directly. Experiment 2.3 was thus designed to test the influence of outcome base rate on the redundancy effect under a 50% intermittent pairing schedule for both redundant cues within the hormone change task.

Method

Participants

One hundred and twelve participants from Prolific Academic were recruited for this experiment. Upon signing up for the study, participants were randomly assigned to either the 0% base rate condition in which hormone levels never increased in the absence of a medicinal cue or the 50% base rate condition in which hormone levels increased on half of occasions even when no medicine was taken. Applying the exclusion criteria resulted in the removal 12 participants, leaving 50 participants in both the 0% group (20 females, mean age=29.86, *SD*=9.08) and the 50% group (21 females, mean age=28.50, *SD*=7.34).

Design and Procedure

Table 2.4

Design of Experiment 2.3

Training	Likelihood ratings test	Forced choice test
Blocking: A+ AX+ I+/– IZ+/–	X Y A B C I Z K L	X vs. Y Z vs. Y
	no cue	X vs. Z
Relative Validity: BY– CY+		
Fillers: K– L–		
No cue: 0% or 50%		

Note. Letters represent cues, randomly assigned to different medicine names. "+" represents hormone increase and "–" represents no hormone change. The outcome never occurred on no cue trials in the 0% group and followed 50% of no cue trials in the 50% group.

The design of Experiment 2.3 is summarised in Table 2.4. On the basis of Experiment 2.1, fillers were included to control the base rate for all cue-present trials at 50% and a 'no cue' trial type was introduced to adjust the outcome probability in the absence of any cues to be either lower than or equal to the outcome rate following the intermittently trained blocked cue, Z. The 'no cue' trials appeared 4 times per block. In the 0% group, the outcome rate in the absence of any cues was substantially lower than that for the intermittently trained blocking procedure (the outcome never occurred without a cue present). The partially reinforced blocking cue, I, would thus provide useful information about the outcome above and beyond the outcome predicted by no cue and compete effectively with Z. The 50% group, on the other hand, implemented a blocking procedure and a 'no cue' trial type that were matched precisely for outcome rate. In this regard, cue I would predict outcome

presence to the same extent as is predicted on ‘no cue’ trials and undergo weaker competition with Z. If the lack of a Z > Y effect in the previous experiments is due to the blocking and blocked cues failing to convey information about an increase in the probability of the outcome relative to base rate, then this base-rate manipulation should influence the relative assessment of Z and Y.

Results

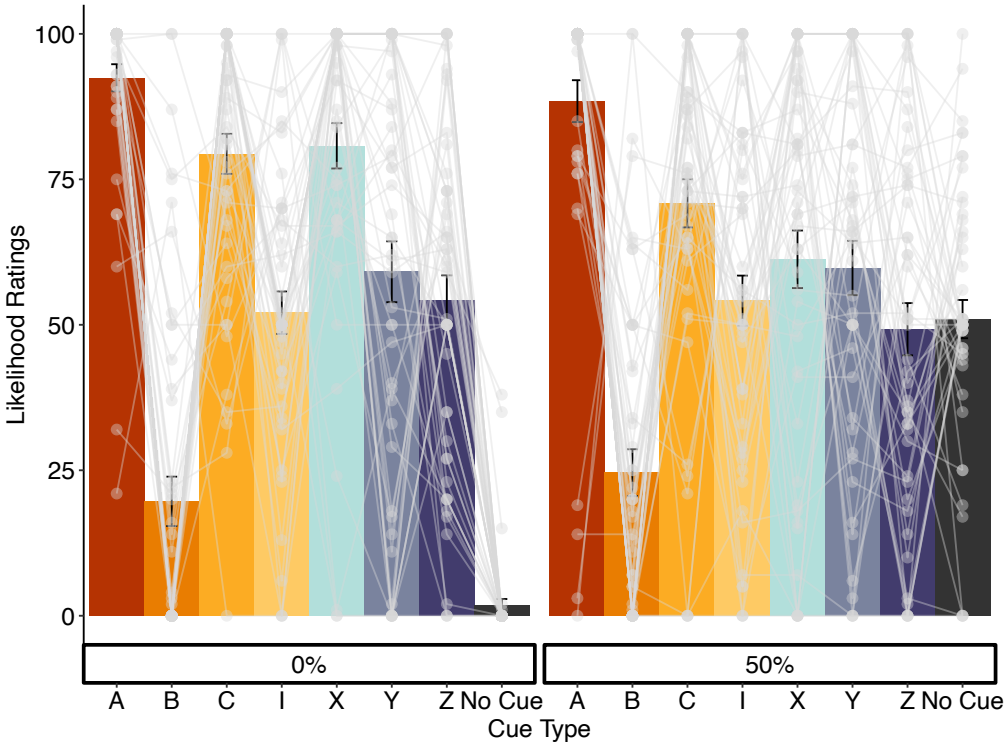
Training

Learning proceeded smoothly across training blocks with the 0% group achieving 85.73% accuracy and the 50% group achieving 89.57% accuracy on cue-present trials with a predictable outcome in the last half of training (please see supplementary materials for training graphs). A 2x(6)x(10) mixed model ANOVA with group (0% vs. 50%) as the between-subjects factor, and trial type and block (1-10) as within-subjects factors revealed a significant quadratic trend for both the 0% group, $F(1,49)=22.22$, $p<.001$, $\eta_p^2=.312$, and the 50% group, $F(1,49)=54.19$, $p<.001$, $\eta_p^2=.525$. Although the 50% group appears to have reached asymptote earlier than the 0% group, this faster learning was not statistically significant, $F(1,98)=1.92$, $p=.169$, $\eta_p^2=.019$.

Likelihood Ratings

Figure 2.6

Mean likelihood ratings on the ratings test in the 0% group (left) and the 50% group (right) of Experiment 2.3. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.



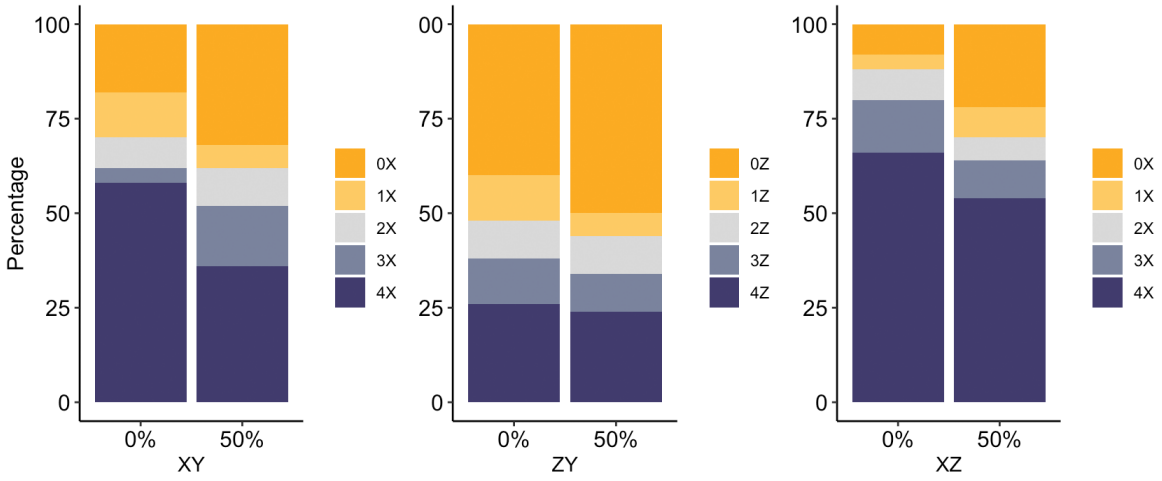
Likelihood ratings for the 0% group and the 50% group are illustrated in Figure 2.6. To assess the effect of base rate on learning about redundant cues, likelihood ratings were compared between four pairs of critical cues: X and Y, Z and Y, A and X, and I and Z. Each pair was analysed using a 2x(2) mixed measures ANOVA with group (0% vs. 50%) as the between-subjects factor and cue type as the within-subjects factor. Results revealed significantly higher ratings for X over Y averaged across groups, $F(1,98)=7.61$, $p=.007$, $\eta_p^2=.072$, and this distinctive pattern for the standard redundancy effect was significantly stronger with 0% base rate than with 50% base rate, $F(1,98)=5.76$, $p=.018$, $\eta_p^2=.056$. Subsequent paired samples t-tests revealed that the redundancy effect was significant in the

0% group, $t(49)=4.08$, $p<.001$, $d=.577$, but not in the 50% group, $t(49)=.23$, $p=.818$, $d=.033$, as supported by the Bayesian binomial test that the ratings for X and Y were equivalent, $BF_{10}=.158$. Further independent samples t-test revealed that X was judged as a significantly more likely cause in the 0% group than in the 50% group, $t(98)=3.10$, $p=.003$, $d=.619$. For redundant cues trained under 50% partial reinforcement, ratings for Z were slightly lower than for Y but not significantly so, $F(1,98)=2.91$, $p=.091$, $\eta_p^2=.029$. This result was consistent across groups, as there was no interaction, $F(1,98)=.39$, $p=.534$, $\eta_p^2=.004$, and no difference between Y and Z in either group individually; 0% group, $t(49)=.71$, $p=.480$, $d=.101$, nor the 50% group, $t(49)=1.79$, $p=.080$, $d=.253$. Further, there were significantly higher likelihood ratings to A than to X averaged across groups, $F(1,98)=34.42$, $p<.001$, $\eta_p^2=.260$, and significantly higher ratings to A and X on average in the 0% group than in the 50% group, $F(1,98)=7.67$, $p=.007$, $\eta_p^2=.073$. The higher ratings to A than to X were significantly more pronounced in the 0% group than in the 50% group, $F(1,98)=5.46$, $p=.022$, $\eta_p^2=.053$. Additional paired samples t-test revealed that X was rated as a significantly less likely cause than A in both the 0% group, $t(49)=2.73$, $p=.009$, $d=.386$, and the 50% group, $t(49)=5.38$, $p<.001$, $d=.761$. In contrast, performing the same analysis on cues I and Z, there was no main effect of cue, $F(1,98)=.09$, $p=.765$, $\eta_p^2<.001$, no interaction with group, $F(1,98)=.60$, $p=.442$, $\eta_p^2=.006$, and no difference between I and Z in either group; 0% group, $t(49)=.37$, $p=.715$, $d=.052$, nor the 50% group, $t(49)=.70$, $p=.487$, $d=.099$.

Forced Choice

Figure 2.7

Mean percentage of choosing the first cue from the XY pair, the ZY pair and the XZ pair on the forced choice test in the 0% group and the 50% group of Experiment 2.3. Higher percentage indicates higher likelihood that a given cue is chosen as a cause.



Results on the forced choice test are depicted in Figure 2.7. Sign tests revealed that X was selected more frequently as the more likely cause than Y in the 0% group, p=.013, but was not in the 50% group, p=.186, in line with the pattern of redundancy effect indexed by likelihood ratings. Percentage of choices for Y was slightly higher than that for Z, but did not reach significance in either the 0% group, p=.186, or the 50% group, p=.068. X was chosen with greater probability than Z in both the 0% group, p<.001, and the 50% group, p=.009, consistent with the ratings test.

Discussion

Experiment 2.3 revealed further evidence for the effect of outcome base rate on the strength of the redundancy effect. The design invoked a 'no cue' trial type to explicitly show the base rate probability of outcome occurrence. When the outcome rate in the absence of a putative cause was 0%, X was regarded as being a more likely cause, compared to when the

base rate was raised to 50%. As a consequence, there was a large redundancy effect (X>Y) in the 0% group but no redundancy effect (X=Y) in the 50% group. This finding does not support the previously discussed role of context in modulating learning about uncertain cues. If outcome base rate as signaled by the common context were to play a role, the higher predictive utility of the blocking cue relative to the explicitly shown 0% base rate should engender stronger blocking of learning about the blocked cue X and hence a diminished redundancy effect. Furthermore, to compensate for the ambiguous causal status, a number of authors have proposed that the likelihood judgement of the blocked cue should increase proportionately with the underlying base rate (Livesey et al., 2013; Jones et al., 2019). The group difference is at odds with this hypothesis.

The reduced ratings for X in the 50% group seem to be a consequence of increased global uncertainty. That is, although the outcome occurred more frequently on 'no cue' trials in the 50% group, the outcome was also more uncertain than that in the 0% group. The increased uncertainty may have led to a specific decline in likelihood judgment for the causally ambiguous blocked cue. Alternatively, the likelihood judgment for X in the 0% group could have been augmented from a particularly strong within-compound association with A which signaled a larger enhancement in outcome probability relative to no cue, compared to A in the 50% group. In principle, the probabilistically trained blocked cue Z should be more sensitive to the group difference in base rate, as a 50% base-rate is equivalent to the outcome probability in the presence of Z, whereas 0% base-rate is noticeably lower. However, the group manipulation appeared to have little effect on judgments of Z either in absolute terms or relative to judgments of cue Y.

Experiment 2.4

Experiments 2.1, 2.2, and 2.3 found evidence of the redundancy effect only when the blocked cue was followed by the outcome on 100% of its presentations but failed to do so

when its absolute relationship with the outcome was weakened to 50%. In all of these previous conditions, the presence of BY and CY trials led to Y being paired with the outcome on half of the occasions yet these trials were nonetheless resolvable through discrimination learning. *Outcome predictability* was high on these trials. Here, we use the term *outcome predictability* specifically to refer to the accuracy with which an outcome can, *in principle*, be predicted with certainty for a given trial type based on all cues presented. This term is different from absolute relationship which refers to the probability that an outcome will follow a particular cue, irrespective of the other cues with which it appears. For instance, Y has an imperfect absolute relationship with the outcome because p(O|Y) = 0.5, but Y is experienced on trials with perfect outcome predictability because p(O|BY) = 0 and p(O|CY) = 1. This was not the case for the partially reinforced blocked cue whose outcome was never fully predictable but determined randomly. That is, Z has both imperfect absolute relationship, p(O|Z) = 0.5, and imperfect outcome predictability, p(O|IZ) = 0.5. One way of resolving the doubt is thus to match the blocked and uncorrelated cues more closely in term of the uncertainty brought about by the probabilistic outcome. Experiment 4 was designed with this aim in mind.

Method

Design and Procedure

Table 2.5

The design of Experiment 2.4

Training	Likelihood ratings test	Forced choice test
Blocking: A+ AX+ I80+ IZ80+ Relative Validity: BY– CY80+ Fillers: K– L–	X Y A B C I Z K L	X vs. Y Z vs. Y X vs. Z

Note. Letters represent cues, randomly assigned to different medicine names. "+" represents hormone increase and "–" represents no hormone change. 80+ represents a hormone increase on 80% of trials (and no hormone change on the other 20%)

As illustrated in Table 2.5, Experiment 2.4 followed both the blocked cue Z and the uncorrelated cue Y with a probabilistic outcome in a way that the two critical cues were both experienced with some degree of aleatory uncertainty. The design and procedure were otherwise the same as per Experiment 2.1.

Participants

Fifty-seven participants from Prolific Academic were recruited for this online experiment in exchange for monetary reimbursement. Three were removed from analyses on the basis of the exclusion criteria detailed above, leaving 54 in the final sample (24 females, mean age=27.52, *SD*=6.92).

Results

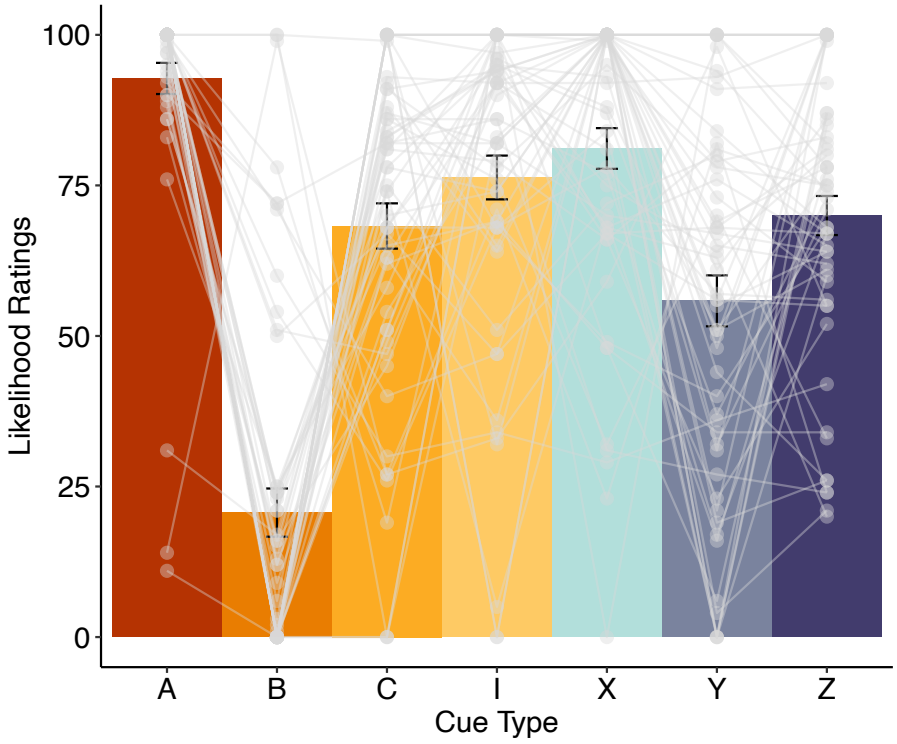
Training

Accuracy on trials with a predictable outcome improved rapidly across training blocks with participants accurately predicting 86.33% of the trials in the last half (please see supplementary materials for training graphs). A (5x10) within-subjects ANOVA with trial type and block (1-10) as independent variables and prediction accuracy as the dependent variable revealed a significant quadratic trend, $F(1,52)=54.04$, $p<.001$, $\eta_p^2=.505$, supporting a rapid increase in accuracy towards ceiling.

Likelihood Ratings

Figure 2.8

Mean likelihood ratings on the ratings test of Experiment 2.4. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

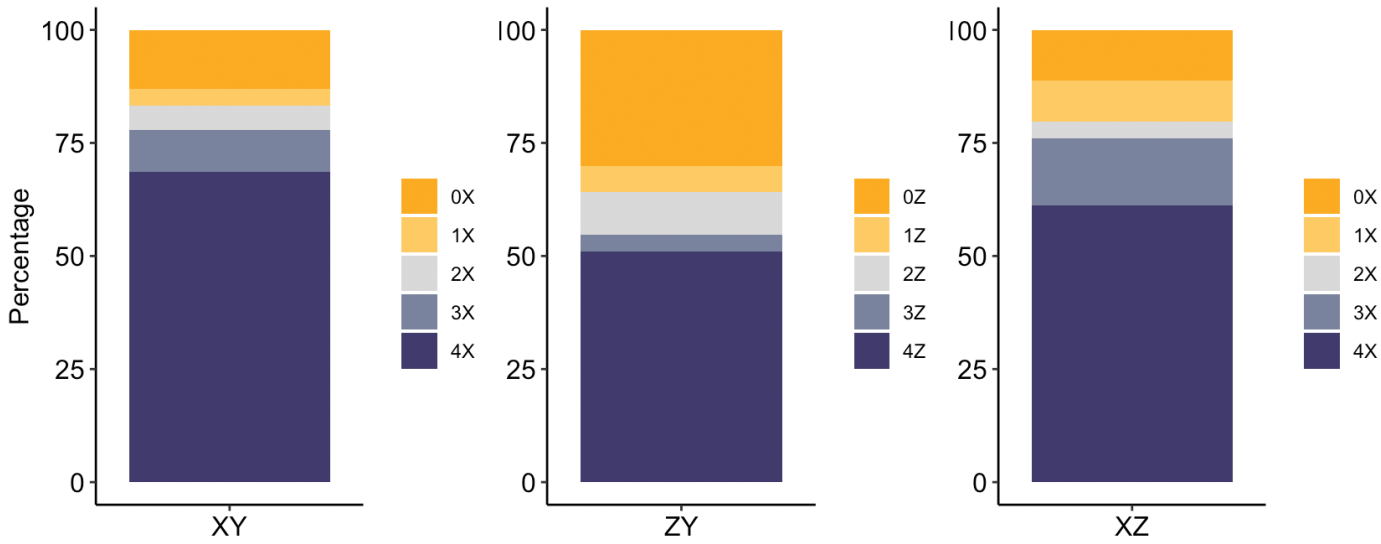


Performance on the likelihood ratings test is illustrated in Figure 2.8. The pairwise comparisons of most interest were between X and Y, Z and Y, A and X, and I and Z. Paired samples t-test for each comparison revealed significantly higher likelihood ratings to X over Y, $t(53)=4.71$, $p<.001$, $d=.641$, indicating the presence of the standard redundancy effect. Interestingly, this time ratings for Z were also significantly higher than for Y, $t(53)=2.95$, $p=.005$, $d=.401$. Ratings for X were significantly higher than for Z , $t(53)=3.12$, $p=.003$, $d=.425$. Moreover, ratings to A were significantly higher than those to X, $t(53)=3.24$, $p=.002$, $d=.441$. Although inspection of Figure 2.8 indicates slightly higher ratings for I than for Z, this difference did not reach significance, $t(53)=1.25$, $p=.216$, $d=.170$.

Forced Choice

Figure 2.9

Mean percentage of choosing the first cue from the XY pair, the ZY pair and the XZ pair on the forced choice test in Experiment 2.4. Higher percentage indicates higher likelihood that a given cue is chosen as a cause.



Results on the forced choice test as shown in Figure 2.9 were analysed through non-parametric sign tests. There was a higher probability of choosing X over Y as the more likely cause, $p<.001$, confirming the redundancy effect in the standard condition. X also had a higher probability of being chosen as the more likely cause than Z, $p<.001$. Although choice probability was numerically higher for Z than for Y, this difference was not statistically significant, $p=.126$.

Discussion

The results of Experiments 2.1-2.3 made us question the extent to which a design in which the blocked cue was imbued with some degree of aleatory uncertainty can produce the redundancy effect. The standard redundancy effect is elicited when the uncorrelated cue is paired with the outcome 50% of the time but the blocked cue consistently followed by the

outcome on 100% of its occurrences. Uengoer and colleagues' (2013) demonstration of an enduring redundancy effect after reducing the outcome probability of the blocked cue to 50% suggests that the characteristic learning difference may be reproducible regardless of whether the redundant cues hold probabilistic or deterministic relationships with the outcome. That is, changes in absolute cue-outcome relationship and outcome predictability brought about by the change in training schedule did not seem to be of critical relevance to the effect in their study. Our results across Experiments 2.1-2.3 clearly contradict this finding. However, Experiment 2.4 found for the first time in this series evidence of the redundancy effect when both redundant cues were trained in a probabilistic setting. Specifically, this experiment found higher likelihood judgment for the blocked cue Z over the uncorrelated cue Y in a design where outcome unpredictability was introduced to both blocking and relative validity. It should be noted that this design did not strictly balance Z and Y in terms of their absolute cue-outcome relationship (i.e. outcome probability) and cannot be taken as direct support for Uengoer et al.'s original demonstration. However, we have at least found evidence of the redundancy effect under conditions in which the occurrence of the outcome is uncertain. The presence of the effect is not merely a result of the outcome following the blocked cue being entirely predictable, something that could not be ruled on the basis of Experiments 2.1-2.3.

Models of Associative Learning

Research into the redundancy effect was initially motivated by its theoretical significance for models of associative learning and continues to revolve around the key theme of testing existing theories and developing new ones. The following section will describe four models that have received most attention in the literature.

The Rescorla-Wagner Model

Rescorla and Wagner's (1972) model is one of the pioneering mechanistic models developed on the basis of cue competition effects including blocking and relative validity. The model posits that learning is proportional to a *summed error* term, defined as the discrepancy between the summed associative strength of all co-occurring cues V_{summed} and the asymptotic associative strength supportable by an outcome λ. This summed error is then multiplied by the salience of the cue α_A and the salience of the outcome β to yield the associative change for cue A on a given trial ΔV_A, as captured by Equation (1).

$$\Delta V_A = \alpha_A \, . \, \beta \, . \, (\lambda - V_{summed}) \qquad (1)$$

Applied to the redundancy effect, the Rescorla-Wager model predicts that X will accrue restricted associative strength because the outcome following the AX compound is predicted by the more strongly established predictor A and there is thus reduced summed error to facilitate further learning about X. In the case of Y, although its associative strength will alternate between increasing on reinforced CY trials and decreasing on non-reinforced BY trials, the model predicts that a portion of the excitatory associative strength gained on CY will be protected from extinction on BY by B acquiring some inhibitory associative strength. This leaves Y with a moderately positive associative value in the end. The finding of higher associative strength for X than for Y characteristic of the redundancy effect was therefore originally presented as a serious challenge to the Rescorla-Wagner model (Pearce et al., 2012).

The Rescorla-Wagner Model with a Common Element

Vogel and Wagner (2017) postulated a variation of the Rescorla-Wagner theory which takes into the account the role of an overlapping component assumed to be shared by all training and testing trials. This modified version holds that the associative strength of any cue or cue constellation is a combined value of its unique and shared components. This common

element approach reconciles the Rescorla-Wagner model with the redundancy effect under certain circumstances. However, it still falls short when there are predominantly outcome-absent trials (Vogel & Wagner, 2017), when the blocking procedure is trained in a staged manner (Uengoer et al., 2020), or when the training length is extended (Uengoer et al., 2019).

The Mackintosh Model

Rather than using a summed error term to account for cue competition effects, other theories have ascribed learning to the associability of the cue, or how well the cue predicts the outcome. According to these models, the alpha value of a cue is not fixed, but varies depending on a cue's predictive utility relative to concomitant others. Specifically, the associability of a cue increases on trials where the cue is better able to predict the outcome, and decreases if there are more predictive companions. This notion of a variable alpha was used in Mackintosh's (1975) classic theory of selective attention. In its original form, the model used an individual error term, that is, the strengthening of an association from a cue to an outcome is proportional to the discrepancy between the associative strength of the individual cue V_A and the asymptotic associative strength λ as given by Equation (2). The change in associative strength is then modulated by the associability of cue A. Mackintosh's original theory did not specify a precise formula for alpha change, however Le Pelley (2004) proposed Equation (3), where V_O is the associative strength of the other cue(s) present and θ is a free parameter that determines the rate of associability change. In a nutshell, the associative strengths of accompanying cues contribute to cue competition through their influence on the target cue's associability, not their involvement in error calculation.

$$\Delta V_A = \alpha_A \, . \, \beta \, . \, (\lambda - V_A) \qquad (2)$$

$$\Delta \alpha_A = \theta . (\, | \, \lambda - V_O \, | \, - \, | \, \lambda - V_A \, | \,) \qquad (3)$$

Mackintosh's model predicts a drop in associability for X and Y at a similar rate as they are both less predictive than the cues accompanying them. However, the outcome-absent BY trials intermittently reduce the associative strength of Y whereas X is consistently paired with the outcome. This difference in training schedule ultimately leads to a weaker association being formed between Y and the outcome than between X and the outcome even though both cues suffer from a similar decline in associability.

The Uengoer Model

Mackintosh's model incorporates the individual error learning algorithm with an associability change mechanism to accommodate the redundancy effect, however, Uengoer et al. (2020) argued that one limitation of the model was that it relies on a restricted range of parameters.[5] To overcome this limitation, Uengoer et al. (2020) devised a simpler way of calculating the change in associability on a trial-by-trial basis, while retaining the individual error learning rule. This model similarly states that the associability of a cue is a function of its predictiveness relative to its associates. A cue's associability increases where it is better at predicting the outcome than others according to Equation (4a), and deceases where it is less predictive according to Equation (4b).

$$\Delta\alpha = \theta.(1 - \alpha) \qquad (4a)$$

$$\Delta\alpha = \theta.(0 - \alpha) \qquad (4b)$$

The Uengoer model has the particular advantage of accounting for the redundancy effect with a 50% probabilistically trained blocked cue independent of the parameters selected. The model predicts a continuous decline in associability for Y throughout training because Y is always less predictive than B and C. The associability of X reduces on reinforced AX trials where A is a better predictor of the presence of the outcome, but increases on non-reinforced AX trials where A becomes a poorer predictor of the absence of the outcome. The net consequence of these variations in associability and the individual error

[5]In light of the findings in this Chapter, especially Experiment 2.1, whether this can now be considered an advantage is highly questionable.

learning rule is therefore the higher associative strength for X over Y, a result that other candidate models struggle to explain.

Summary

The four models under primary consideration in the present paper are the Rescorla-Wagner model (1972), the modified Rescorla-Wagner model incorporating a common element (Vogel & Wagner, 2017), the Mackintosh model (1975), and the Uengoer model (2020). Among these candidate models, Uengoer's model appears to be the most successful so far, as it plausibly explains all demonstrations of the redundancy effect and related findings in the available literature. Rescorla and Wagner's model predicts higher associative strength for Y over X, a pattern opposite to that observed in the redundancy effect. Its later version revised by Vogel and Wagner (2017) predicts a transient redundancy effect depending on training length (Uengoer et al., 2019) but fails to predict the effect when the overall outcome rate is low (Vogel & Wagner, 2017) and when the blocking procedure is trained in a staged rather than an interspersed manner (Uengoer et al., 2020). Mackintosh's theory of selective attention anticipates the effect only with a strictly selected set of parameters if using staged training for the blocking procedure (Uengoer et al., 2020). It is clear that each of these models presents its own drawbacks in explaining the redundancy effect.

The plausibility of the main mechanisms employed by the four models will be tested both in isolation and in combination. The influential Rescorla-Wagner model has formed a cornerstone in the development of associative learning theories. Its application to the redundancy effect is however non-viable without additional assumptions being made about the shared component across training and test trials. It is also common practice to represent cues as constellations of overlapping features (McLaren & Mackintosh, 2000, 2002; Harris, 2006; Harris & Livesey, 2010). For these reasons, we will take a common cue approach to

examine the summed error learning rules. Moreover, the error-correction algorithms and alpha change mechanisms are not mutually exclusive but are indeed both assumed in attentional models. It is unclear if employing both is more advantageous over one. Some mechanisms may be useful for certain effects but have limitations for others. It is therefore important to examine the generalisability of these mechanisms to capture the effects in context.

Computational Modelling

Computer simulations were run to enable a direct comparison of the empirically collected data with predictions derived from different learning theories. All simulations in this section were programmed in R using RStudio. The models under consideration in this section consist of either the separable or the summed error learning algorithm, each without attention change, with the Mackintosh attention change algorithm, and with the Uengoer attention change algorithm, thus yielding 6 models in total. In all 6 models, a common element assumption was embedded in the simulation where the context is regarded as a shared cue that is present on all training trials. The inclusion of a common element provides a means for the Rescorla-Wagner model to account for the classic redundancy effect. Although some authors have noted important limitations of this explanation, we include the common cue here because, in principle, we feel it is important for the commonalities between trial types to be represented in the model in some way. (See Supplementary materials for some additional discussion and post-hoc analyses concerning the implications of using a common cue approach).

The purpose of this modeling was to understand 1) whether, in principle, these models can provide a reasonable account of the judgments of ambiguous cues, 2) which model provides the best fit, 3) how the predictions of the key models vary with variations in key parameters. Aims 1 and 2 were addressed by fitting the model to training and cue choice test

data using the simplex algorithm and then examining predictions for the ratings phase generated under the best-fitting parameters. Aim 3 was addressed by varying several parameters systemically (using a simple grid-search method) and examining how predicted differences between the ambiguous cues vary across these parameter values.

Parameter Fitting

We first conducted parameter fitting to examine model predictions under optimal parameter values. Across the current series of 4 experiments, and for all models, the precise trial orders experienced by participants were used to run the simulated experiments. We used two sources of data to fit each model; training trials from second block onwards and all cue choice test trials. Note that this means the best-fitting parameters yield predictions which are independent of the ratings provided by participants. We decided to do this partly for convenience (unlike continuous causal ratings, the training and cue choice test provide discrete choices for which there are widely used decision rules (e.g. Don et al., 2019) that can yield likelihood estimates in a principled way) but also because it provides a more rigorous test of the model's predictive capabilities, since it assumes the same associative weights are used to derive responses in each of the three phases.

On each trial, to generate model likelihood predictions, the associative strengths of all cues present (including the common cue) were added together to yield a summed associative weight. From this combined associative strength, a prediction probability, *P*(outcome prediction), was generated for predicting 'increase' using the cumulative density function of a standard normal distribution (with mean=0 and sd=1):

$$P(\text{outcome prediction}) = P(X < C \cdot (V_{summed} - c2)) \qquad (5a)$$

$$C = 3^{c1} - 1 \qquad (5b)$$

Here, V_{summed} is the summed associative strength of cues present on a training trial, and c1 and c2 are scaling parameters that adjust the extent to which the summed associative strength

of trial cues drives prediction of the outcome. The cumulative density function calculates the probability that a normally distributed random variable takes any value that is less than $C \cdot (V_{summed} - c2)$, that is, the total area under the bell curve to left of the scaled associative weight. The probability of choosing 'no change' took the complement of this prediction probability, i.e. 1 - *P*(outcome prediction). This approach is used to provide prediction probabilities for a binary choice, based on a single response strength value from the model (e.g. see Don & Worthy, 2021). Based on participants' response on a given training trial, the corresponding likelihood of generating the same prediction was selected for that trial.

On the cue choice test, each individual cue held a particular associative strength as a result of training in the preceding phase. The associative strengths of the two cues, V1 and V2, within each pair of comparison were pitted against each other. Their competing associative weights were converted into a probability mass function with two discrete variables through the SoftMax rule. The SoftMax is a widely implemented classification strategy that operates on multiple variables including binary outcomes as follows:

$$P(\text{choose cue 1}) = \frac{e^{D*V1}}{e^{D*V1} + e^{D*V2}}$$

$$D=3^{c3} - 1$$

Here, c3 is a scaling parameter that adjusts the probability difference between the two choices on the basis of their associative strengths. The probability of selecting the cue that participants indicated on each test trial was determined by referring to the corresponding outcome in the probability distribution.

The product of all prediction probabilities and choice probabilities was transformed into a negative logarithm. The negative log likelihood function was minimised to maximise the likelihood of observing the empirical data given the most plausible combination of parameters for each model. This minimisation was achieved using the simplex method which estimates parameters through the Nelder-Mead algorithm (Nelder & Mead, 1965). Although

the simplex method operates efficiently in many circumstances, the function may not always return the true minimum but be trapped by local minima in which case repetitions would be necessary. The optimisation procedure was thus run 100 times for reliability. The best fit value was taken as an index of a model's capability to predict the empirical data. However, Akaike information criterion (AIC) and Bayesian information criterion (BIC) were calculated to balance the goodness of fit with the number of free parameters, since the models with no attention change have one fewer free parameter. Models were also evaluated with respect to the practicability of the parameters (e.g. extremely high alpha values near 1 are unlikely as starting parameters).

The starting parameters for the optimisation process was each randomly selected from a uniform distribution bounded within a certain range. Specifically, alpha, beta, theta, and c2 were values between 0 and 1, while c1 and c3 were values between 0 and 5. Although simulations conducted by a number of other authors did not impose restrictions on these parameters in subsequent training trials, an additional bound to confine alpha within the range between 0 and 1 was applied to models enlisting an attention change mechanism to avoid extreme lambda values. Note that the optimisation of scaling parameters is particularly important because it acted as the basis for the systematic variation of alpha, beta, and theta in grid search.

Table 2.6

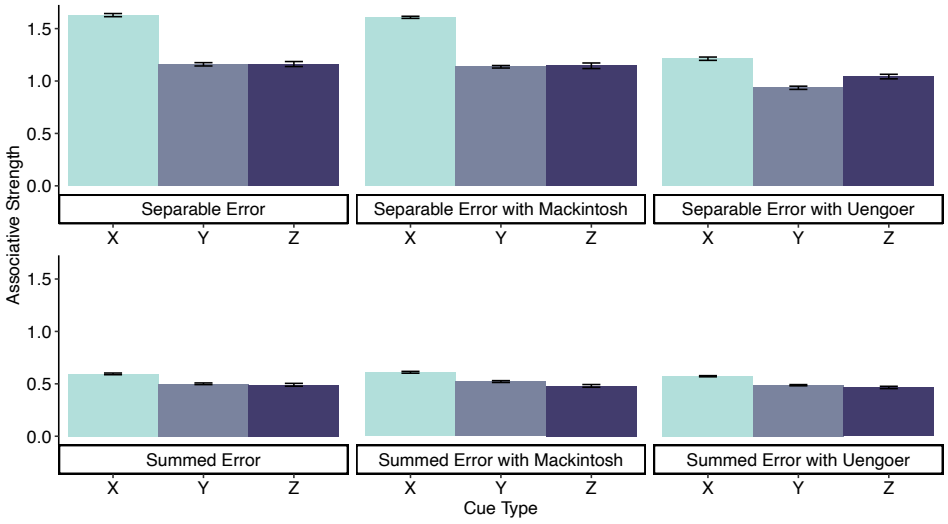
Model fitting for Experiments 2.1-2.4 with training and cue choice data

Experiment	Model		Parameters						Fit statistics		
	Attention	Error term	Alpha	Beta	Theta	c1	c2	c3	- Log L	AIC	BIC
2.1	None	separable	0.245	0.634	-	0.672	1.000	0.990	3304.157	6618.315	6695.108
	None	summed	0.115	0.728	-	1.313	0.459	1.958	2813.505	5637.010	5713.803
	Mackintosh	separable	0.728	0.281	0.214	0.725	1.000	0.897	3269.537	6551.073	6643.225
	Mackintosh	summed	0.837	0.107	0.066	1.301	0.456	2.049	2805.511	5623.021	5715.173
	Uengoer	separable	0.601	0.558	0.379	0.777	0.971	1.463	3202.938	6405.370	6497.522
	Uengoer	summed	0.999	0.094	0.017	1.305	0.456	2.016	2804.969	5621.695	5713.847
2.2	None	separable	0.410	0.405	-	0.533	1.000	0.757	12455.753	24921.507	25009.803
	None	summed	0.307	0.274	-	0.965	0.438	0.955	11997.682	24005.364	24093.661
	Mackintosh	separable	0.999	0.209	0.049	0.520	1.000	0.667	12418.276	24848.545	24954.501
	Mackintosh	summed	0.969	0.102	0.123	0.967	0.431	1.256	11956.012	23924.024	24029.979
	Uengoer	separable	0.682	0.244	0.033	0.545	1.000	0.881	12432.038	24876.075	24982.031
	Uengoer	summed	0.806	0.121	0.014	0.959	0.435	1.012	11969.084	23950.168	24056.123
2.3	None	separable	0.239	0.716	-	0.759	1.000	0.871	11301.624	22613.247	22701.874
	None	summed	0.255	0.346	-	1.236	0.462	1.688	10573.208	21156.416	21245.042
	Mackintosh	separable	0.974	0.223	0.237	0.771	0.996	0.872	11098.175	22208.351	22314.703
	Mackintosh	summed	0.766	0.121	0.021	1.229	0.465	1.727	10550.144	21112.288	21218.640
	Uengoer	separable	0.899	0.289	0.269	0.798	1.000	1.182	11112.615	22237.229	22343.581
	Uengoer	summed	0.925	0.109	0.011	1.228	0.467	1.976	10538.599	21089.198	21195.550
2.4	None	separable	0.251	0.591	-	0.835	1.000	1.074	4396.202	8802.404	8882.605
	None	summed	0.540	0.187	-	1.279	0.456	2.119	3858.386	7726.773	7806.974
	Mackintosh	separable	1.000	0.206	0.289	0.860	1.000	1.226	4273.424	8558.847	8655.089
	Mackintosh	summed	1.000	0.136	0.094	1.226	0.454	2.455	3805.048	7622.096	7718.338
	Uengoer	separable	0.992	0.242	0.331	0.941	0.981	0.840	4211.210	8434.421	8530.663
	Uengoer	summed	1.000	0.129	0.020	1.253	0.456	2.224	3822.684	7657.368	7753.610

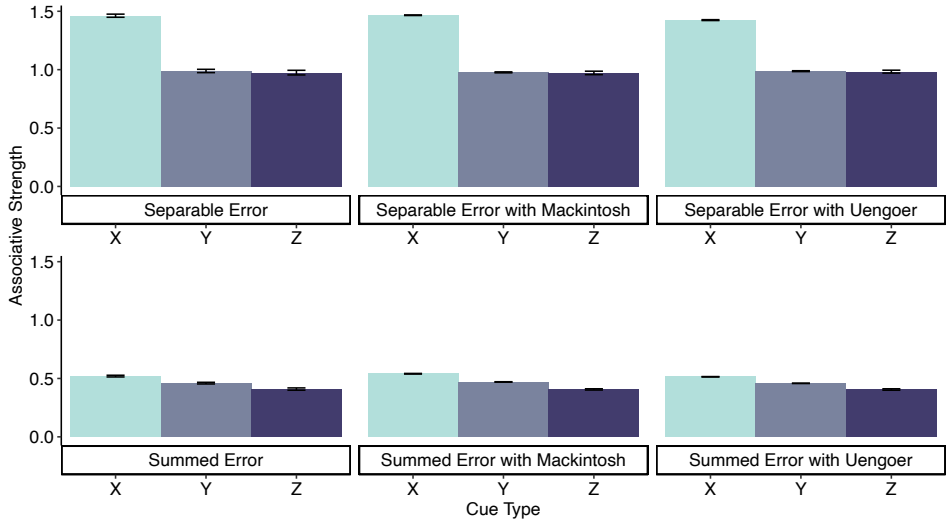
Figure 2.10

Simulated associative strengths for X, Y, and Z under the separable and summed error learning models, each without attention, with Mackintosh attention, and with Uengoer attention across (a) Experiment 2.1, (b) Experiment 2.2, (c) Experiment 2.3, and (d) Experiment 2.4. Error bars indicate standard error of mean (SEM).

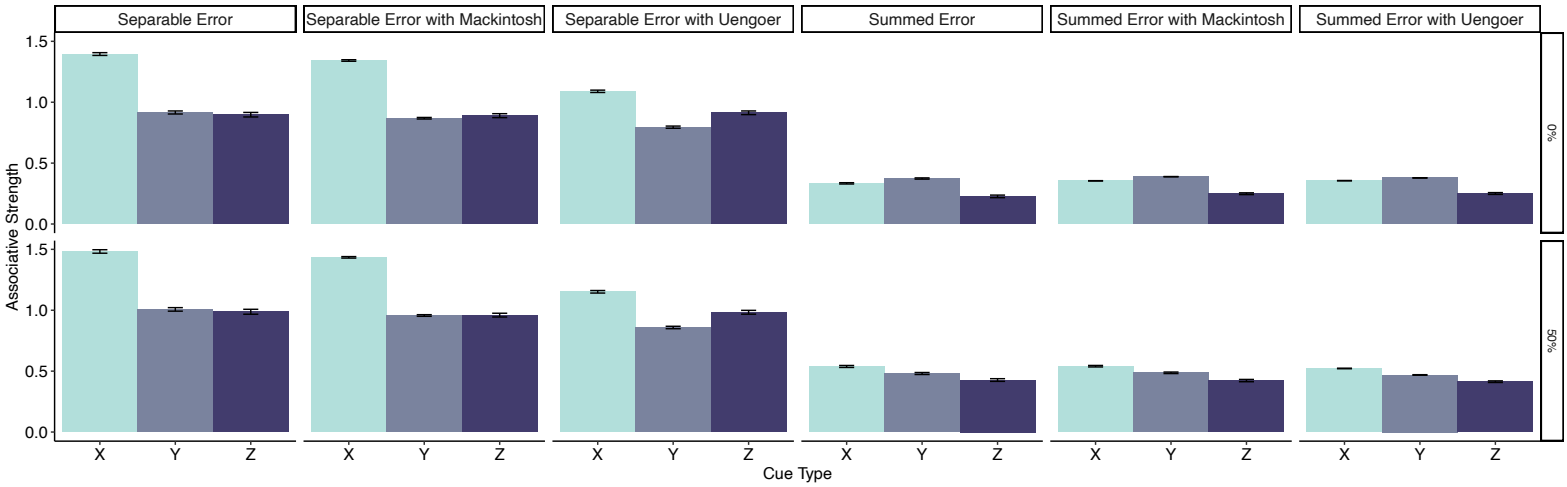
(a)



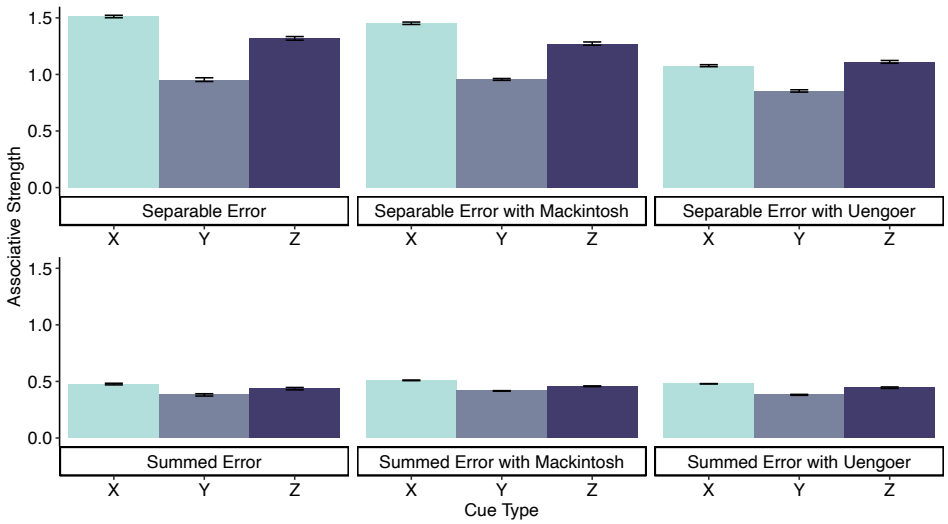
(b)



(c)


Separable Error
Separable Error with Mackintosh
Separable Error with Uengoer
Summed Error
Summed Error with Mackintosh
Summed Error with Uengoer
0%
50%
Associative Strength
1.5
1.0
0.5
0.0
X
Y
Z
Cue Type

(d)

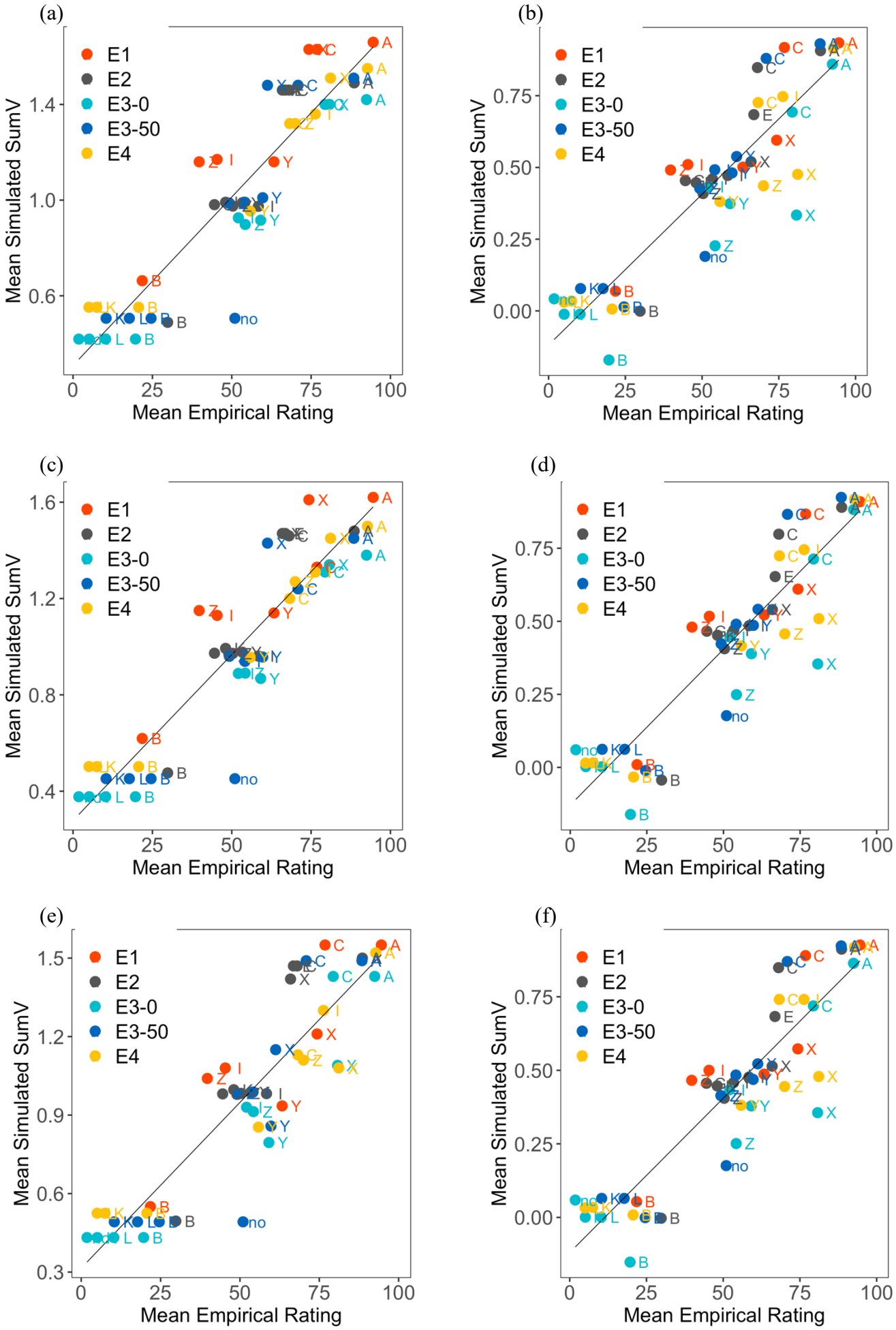

Associative Strength
1.5
1.0
0.5
0.0
Separable Error
Separable Error with Mackintosh
Separable Error with Uengoer
Summed Error
Summed Error with Mackintosh
Summed Error with Uengoer
X
Y
Z
Cue Type

Results from simulations are summarised in Table 2.6 and predicted associative strengths for redundant cues X, Y, and Z under different models are shown in Figure 2.10. The best fit value represents the negative log likelihood of obtaining the empirical data under each model. Across the current series of 4 experiments, the error-correction algorithm based on summed associative strength outperformed that based on the individual associative strengths, irrespective of whether an attention mechanism was incorporated. This is far and away the clearest and most consistent finding, comparing across the six models (see supplementary materials for slight variations in model fitting using training data only and choice data only).

The incorporation of an attention change algorithm into the error correction model slightly improved model fit. Although it adds inherent complexity (and a free parameter) to the learning algorithm, a predictiveness-based attention change algorithm (either that based on Mackintosh, 1975 *or* Uengoer et al. 2020) generally reduced AIC and BIC estimates relative to the equivalent models without attention change. There were exceptions, however; for instance, in Experiment 2.1, BIC was lowest for the summed-error model with no attention change (since a difference in BIC of less than 2 is considered negligible, there was virtually nothing separating the three summed error models). Overall, it is striking that the attention change mechanism provides very modest additional predictive power.

Figure 2.11

Correlational mapping of empirical ratings onto simulated associative strength under (a) Separable error with no attention, (b) Summed error with no attention, (c) Separable error with Mackintosh attention, (d) Summed error with Mackintosh attention, (e) Separable error with Uengoer attention, and (f) Summed error with Uengoer attention. E1-E4 represent Experiments 2.1-2.4 in the current series and individual points represent cues in the ratings test of each experiment.


(a)
(b)
(c)
(d)
(e)
(f)
E1
E2
E3-0
E3-50
E4
Mean Simulated SumV
Mean Empirical Rating

To evaluate whether they produce a sensible estimate of the ordinal pattern observed on the ratings test, simulated ratings need to be derived from associative strengths accumulated for individual cues as well as the common context (i.e. the summed V computed for each trial). In Figure 2.11, the empirical ratings are plotted against simulated predictions under the six models of interest.

For Experiments 2.1 and 2.2, the summed error learning algorithm predicts associative strengths for the three critical cues in descending order (i.e. X>Y>Z). Models based on separable error learning rules do not consistently predict this judgment pattern but predict a sustained redundancy effect under 50% intermittent training for the blocked cue. The pattern of results is more complex with manipulations of outcome base rate in Experiment 2.3. Recall that in this experiment, the redundancy effect (X>Y) was significantly reduced in the 50% base-rate condition compared to the 0% condition. Neither class of error correction model anticipates this group difference. The separable error models tend to predict the redundancy effect irrespective of the outcome rate in the absence of a cue. The summed error models (and the separable error model with Mackintosh attention) predict a trend in the *opposite* direction with a *larger* difference between X and Y in the 50% group. Note that in general summed error models predict this reversed pattern observed more accurately than separable error term models. Lastly, for Experiment 2.4, all six models capture the persistent nature of the redundancy effect when both AX and CY trials were paired probabilistically with the outcome. Overall, the summed error learning algorithm provides a superior basis over the separable error learning algorithm for explaining the judgment pattern of redundant cues, while attention has a modest influence that changes the ordinal trend in minor ways.

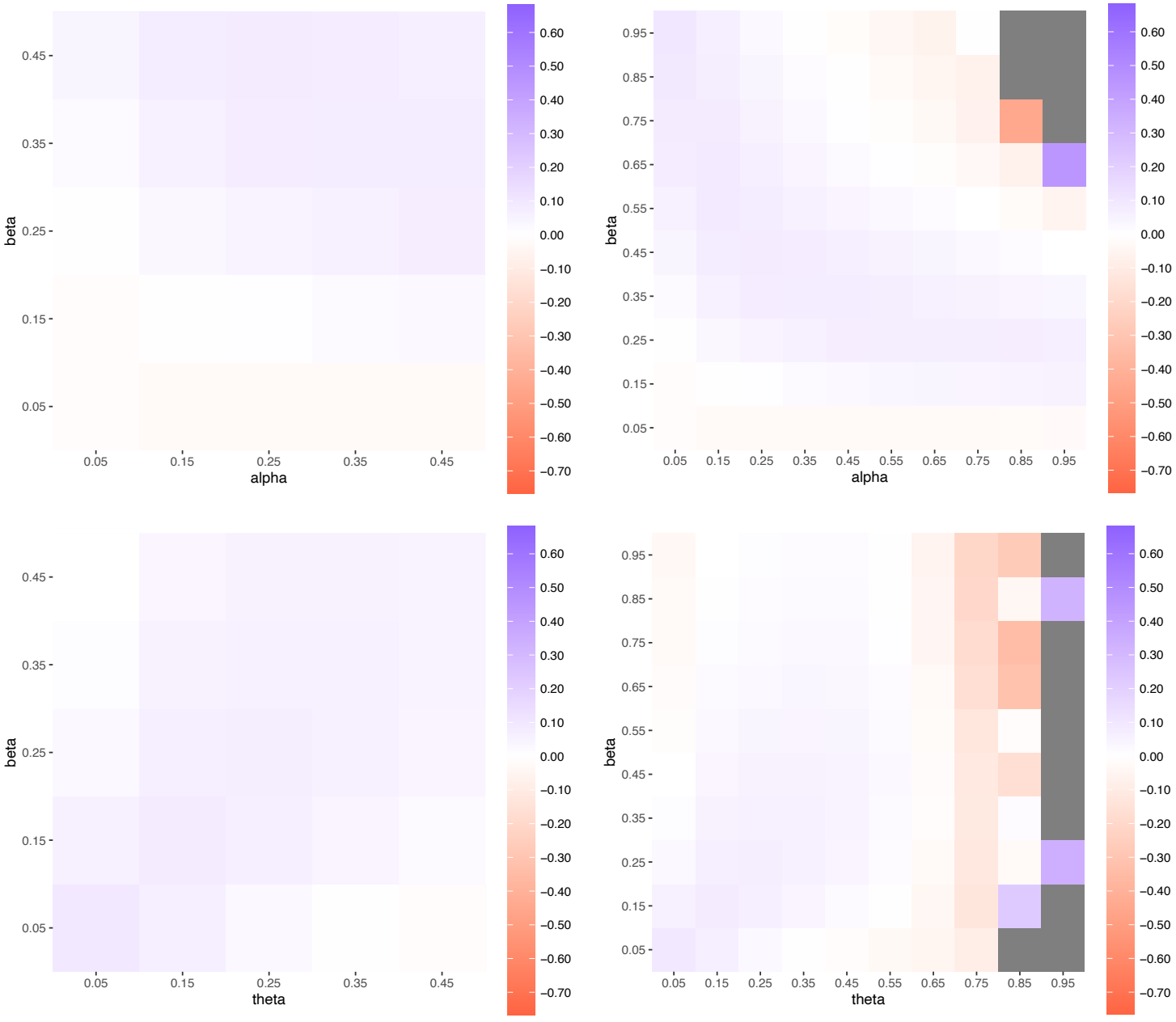
Grid Search

The third aim of the simulations was to examine how the model predictions—especially those relating to relative differences between the key ambiguous cues (X, Y, Z)—vary with variations in key learning parameters. We focus here on the summed error models that use either the Mackintosh attention algorithm or the Uengoer attention algorithm as these models produced the best accounts of the data in the previous section. We conducted an exhaustive search through the alpha-beta and the beta-theta parameter spaces, in each case calculating differences between the X and Y cues and the Z and Y cues. Parameters that were not systematically varied were fixed at their optimal values derived from the model fitting reported above. For each pairwise search, 10 equally sequenced values between the lower bound 0.05 and the higher bound 0.95 were considered for each parameter. This resulted in the evaluation of a total of 100 parameter combinations for each pairwise comparison. For each experiment and each parameter combination, the difference in associative strength between X and Y, and between Z and Y, was calculated after training the model on all trials presented during the learning phase.

Figure 2.12

An example of grid search through the alpha-beta and beta-theta parameter spaces under the summed error term model with (a) Mackintosh attention and (b) Uengoer attention for the standard redundancy effect (X > Y), in Experiment 2.1. Values for one parameter are spaced out along the x-axis and the other along the y-axis. Note that purple areas represent positive differences, orange area represent negative differences, and greyed-out areas represent unrealistically extreme differences.

(a)



(b)

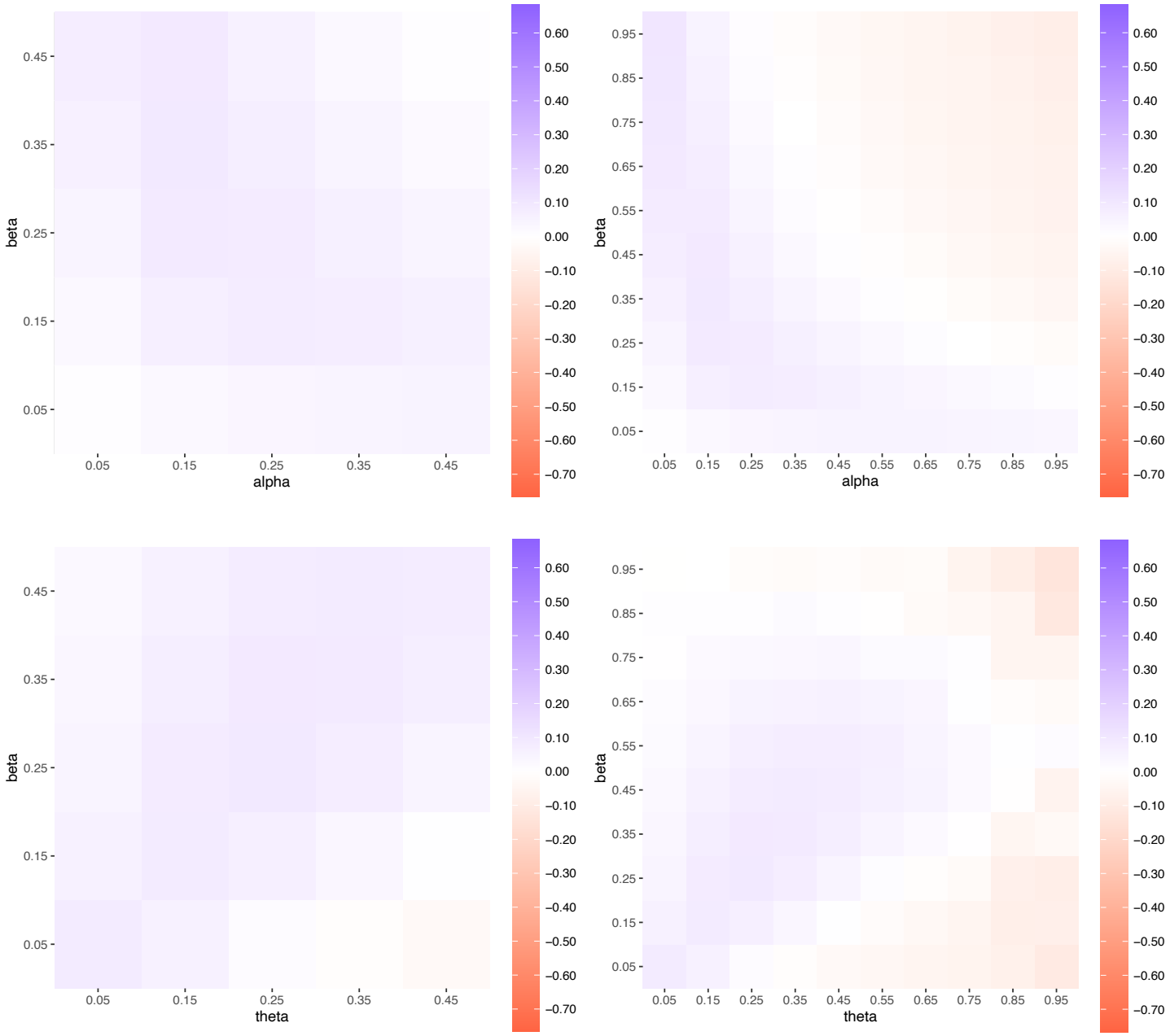
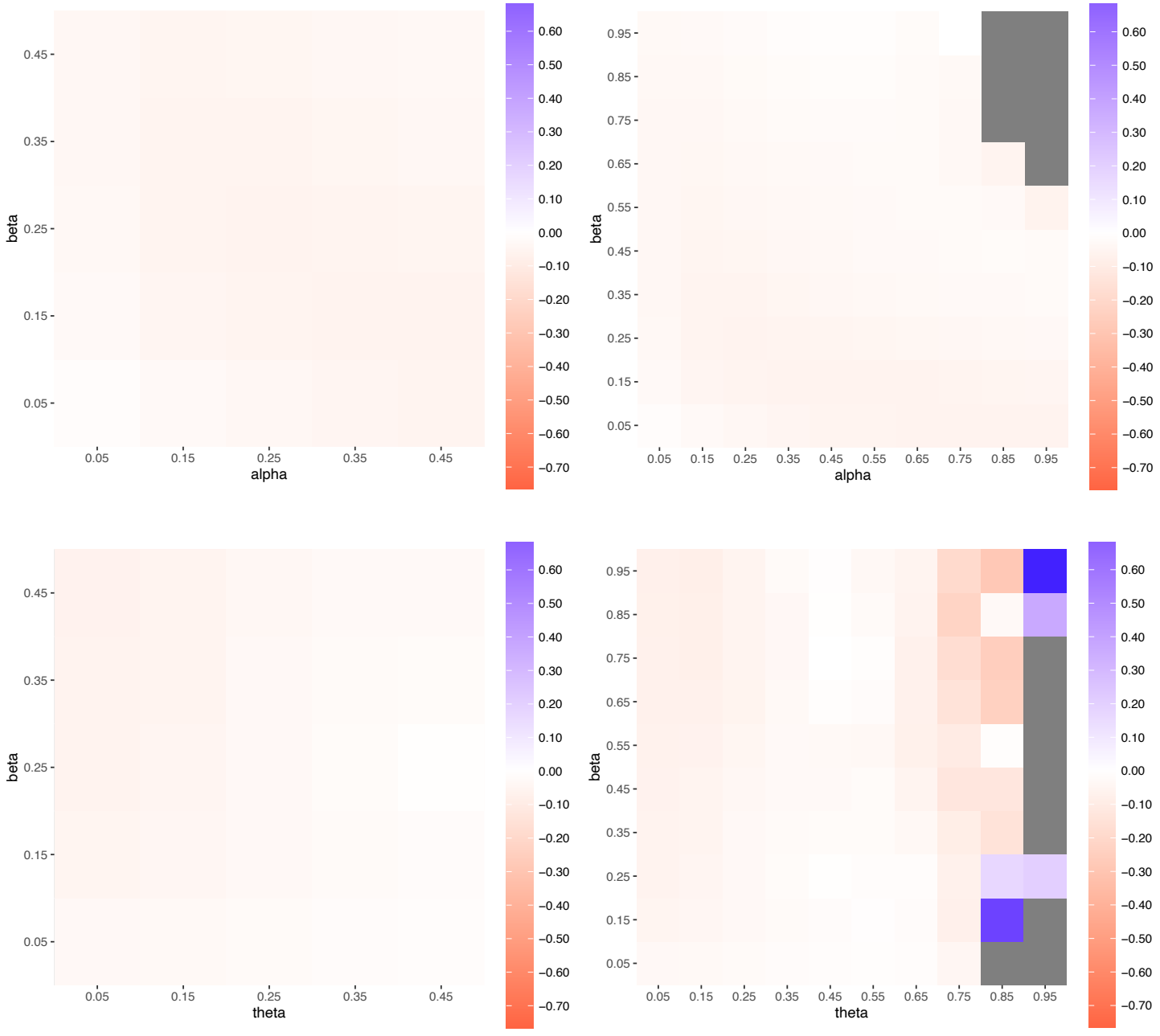

beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

Figure 2.13

An example of grid search through the alpha-beta and beta-theta parameter spaces under the summed error term model with (a) Mackintosh attention and (b) Uengoer attention for the redundancy effect with a partially reinforced blocked cue (Z>Y) in Experiment 2.1. Values for one parameter are spaced out along the x-axis and the other along the y-axis. Note that purple areas represent positive differences, orange area represent negative differences, and greyed-out areas represent unrealistically extreme differences.

(a)



(b)

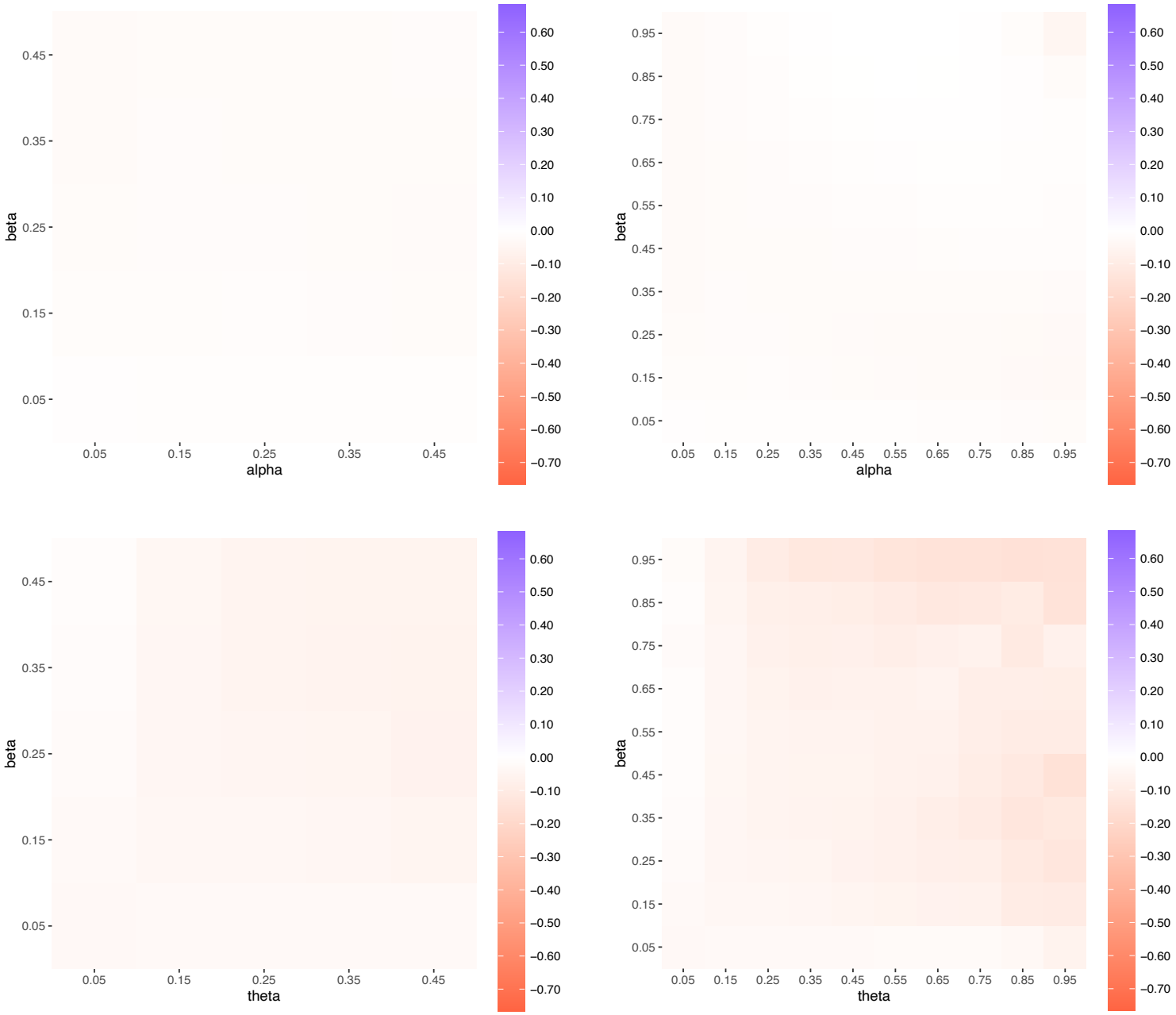


Figure 2.12 and 2.13 illustrate the simulated variations in the standard redundancy effect and the redundancy effect with a partially reinforced blocked cue respectively with different alpha, beta, and theta values as heat maps. I present the results for Experiment 2.1 here, but the results are similar across all four experiments. For each key parameter pair, intersections of two parameter values are represented as grids with darker colours signifying

more extreme differences between the critical cues of interest. Unrealistically large positive and negative associative strength differences can occur when parameter combinations yield very fast learning. As such, any differences in associative strength greater than 1 or less than -1 are greyed out. Speaking of the broad capability to predict the hallmark result, the Mackintosh attention is more reliant on the chosen parameter span than is the Uengoer attention. For example, controlling alpha at its optimal value, the Mackintosh attention predicts the X>Y and Z<Y results observed in Experiment 2.1 when theta value is not too extreme. By contrast, the Uengoer attention predicts the same general trends over much larger variations in the chosen theta. The greater dependence on parameter span implies that the Mackintosh model anticipates learning about ambiguous cues to vary with cue and outcome salience, confining the redundancy effect to carefully defined ranges of parameter values. The Uengoer model expects a more stable redundancy effect with changes in cue and outcome salience. Although sensitivity of the redundancy effect to different circumstances was not directly tested in the current experiments, this is a clear distinction between the two attention models that awaits further empirical support (Please see supplementary materials for the full results from grid search).

Discussion

On the whole, the modelling analyses support associability change as a contributing factor in producing the overall redundancy effect results, in addition to a basic error-correction algorithm. If alpha plays a primary role in the observed learning advantage for X over Y, then it seems reasonable to expect a pattern of attention deployment consistent with this associative difference. That is, more attention should be allocated to X than to Y, which then drives greater learning about the former. For models using separable error terms, relative alpha changes should dominate. The Mackintosh attention with a separable error term predicts greater alpha for X over Y for most experiments except for Experiment 2.2 (see

supplementary materials for more details). Similarly, the Uengoer attention with a separable error term predicts greater alpha for X over Y for at least a portion of training in all experiments, even if they converge on zero eventually. Attention patterns may not align with differences in associative strength for summed error term models as the common error term controls competitive learning. However, the simulated alpha change for the Uengoer attention with a summed error term shows that X gains higher alpha than does Y, across all experiments. On the other hand, if alpha exerts only a minor influence additional to the dominant role of error-correction mechanism in shaping the learning trend of redundant cues, then it is still quite possible that the redundancy effect could be observed in situations where equivalent attention is paid to X and Y, or even where greater attention is allocated to Y than to X. In line with this view, the other three attention models predict a general drop in alpha for redundant cues with X eventually having a lower alpha than Y. This suggests that while alpha makes an extra contribution to predicting the overall results, its role in producing the critical XY difference specifically must not be central.

Influence of Attention on Learning

While attentional explanations are plausible candidates based on theoretical grounds, two empirical studies investigating alpha in the form of overt eye gaze (Jones & Zaksaite, 2018) and covert cue associability (Uengoer et al., 2019) found negligible attention differences between X and Y despite revealing greater causal judgment for the former. Uengoer and colleagues argued that the main driver for the redundancy effect is likely to lie elsewhere, but this lack of attention difference does not preclude attention theories from explaining related findings. In particular, the Mackintosh attention with a separable error term would account for not only a lack of attention difference between X and Y, but also the seemingly odd equal causal judgment for the blocking cue I and the blocked cue Z under the 50% partial reinforcement schedule. They made the further point that these two key findings

cannot be easily reconciled with a common error term model. In contrast, in the current study, summed error models consistently provide better accounts of the data overall and produce an ordinal pattern of associative strengths that matches participant ratings.

In our current simulations, the Mackintosh model predicts higher alpha for X than for Y with a separable error term, but predicts the opposite with a summed error term. That is, rather than generating no perceivable difference between X and Y, simulations of the Mackintosh model yield attention differences between the two critical redundant cues with either form of error-correction rule. Moreover, insofar as the shared element is taken into account, I and Z involved in the 50% reinforcement blocking procedure are expected to undergo comparable associative increment by most summed and separable error models. The only exception is in the condition where the outcome never occurs in the absence of a specific cue (Experiment 2.3, 0% condition), in which case the summed error models predict greater learning about I over Z. The lack of learning difference between I and Z, a result originally reported by Uengoer et al. (2013), is thus not exclusively predicted by the Mackintosh model with a separable error term, but would indeed be explainable by a broad range of models after consideration of the common cue.

Table 2.7

Summary of Pearson correlation coefficients for X vs Y , Z vs Y , and X vs Z differences simulated under the common element model across Experiments 2.1-2.4

	Experiment 2.1	Experiment 2.2	Experiment 2.3	Experiment 2.4
X vs Y	0.333	0.083	0.197	0.153
Z vs Y	0.284	0.059	0.068	0.304
X vs Z	0.203	0.155	0.181	0.392

Role of the Common Cue

In all of the simulations that we consider here, a common cue was used to capture qualities that are common across trials. While the two attention mechanisms provide a modest benefit to model fitting, the inclusion of this common cue seems to play a leading role in predicting a number of key empirical results in the first place. For instance, without it, a summed error term with no attention modulation (e.g. the Rescorla-Wagner model) fails to account for the redundancy effect at all (Pearce et al., 2012). A common cue can directly control model predictions by virtue of its own associations, but can also indirectly influence learning about other cues, especially when using a summed error term. The learning expressed in response to individual cues could thus be highly dependent or show some degree of independence, depending on specifics of the model and the particular contingencies on which it was trained.

To explore this further, we sought to examine interdependencies between the level of learning for cues X, Y, and Z across the models. Table 2.7 reports Pearson correlation coefficients for the simulated predictions of the ratings phase data under the summed error term model. Note that these correlations remain the same for the common element model without attention, with Mackintosh attention, or with Uengeor attention (see supplementary materials for scatter plots). These analyses revealed weak positive correlations overall in the predictions for redundant cues across individual simulated participants in all four experiments. Note that the same set of best-fitting parameters was used for each experiment, and thus the variance in model predictions across runs is a result of the randomised sequences of trials presented to individual participants (these are illustrated in scatterplots in Supplementary Materials). The low degree of inter-relatedness in the model predictions suggest that judgment about one redundant cue has a slight impact on how other redundant cues are treated. This also implies that the common context does not have a uniform

influence over learning about redundant cues. Differential learning observed among X, Y, and Z therefore appears to reflect the independent evaluation of these cues based on experienced contingencies.

Researchers have used a summed error model with a common cue to simulate the redundancy effect. However, one of the criticisms of this approach has been the demonstration that, with extended training, the prediction that X should have a higher associative strength than Y tends to disappear or even reverse (Uengoer et al., 2018). Although we used the precise training sequences to train the models in this study, it remains of interest whether the models continue to make consistent predictions (i.e. X > Y) when the models receive extended training. To examine this, we reran the simulation but repeating the training phase multiple times for each simulated participant. Under the common element model alone, the learning advantage for X over Y is predicted to reduce as training length increases and is predicted to be abolished with further training. However, a reversal of the redundancy effect is not expected even after extending training to 80 blocks. The same transient pattern holds true when the Uengoer attention was added. Interestingly, the X learning bias remains robust when the common element model is combined with the Mackintosh attention even after 80 blocks of training. These results demonstrate partial consistency with Uengor et al. (2018), but the susceptibility of the redundancy effect to training length appears to be related to the attention mechanism employed. This result suggests that the plausibility of the common element model may be further improved by combining a form of alpha change mechanism that tunes attention towards cues with greater predictive utility.

Summary

In summary, computer simulation of error-correction learning algorithms with and without an alpha change mechanism revealed evidence in support of summed error term

models. With the inclusion of a common cue, the basic error-correction principle embodied in the Rescorla-Wagner model serves as an adequate explanation for the redundancy effect and related findings. Integrating a learned predictiveness mechanism such as the theory posited by Mackintosh and Uengoer further improves the model's capability to capture the overall results.

General Discussion

The current research explored learning about the blocked and uncorrelated cues under a range of probabilistic circumstances. In a manner consistent with the blocked cue learning bias (X>Y) observed using the standard (deterministic) design, it was hypothesised that the effect should persist with probabilistic contingencies as long as the blocked cue maintains higher relative informativeness than the uncorrelated cue. This hypothesis assumes that the redundancy effect is solely determined by the informativeness of the critical cues relative to concurrently presented competing cues. The 50% partially reinforced blocking procedure presents a situation in which the blocked cue is equally informative as the blocking cue, while the uncorrelated cue is less informative than cooccurring cues. According to Uengoer et al.'s (2013) initial proposal, this differential informativeness should suffice for a learning difference between the blocked cue and the uncorrelated cue even though both cues signal a 50% chance of outcome occurrence. However, attempts to replicate the redundancy effect under these conditions resulted in a complete reversal of the characteristic differences in likelihood judgment in Experiment 2.1 and a partial reversal in Experiment 2.2. In Experiment 2.3, probabilistic training of the blocked cue abolished the redundancy effect regardless of whether the outcome probability in the absence of a cue was lower than or equal to that for the partially reinforced blocked cue, suggesting that these results cannot rely on how much more predictive utility that the blocked cue provided additional to the context. Taken together, these findings indicate that the learning advantage for the blocked cue over

the uncorrelated cue is not a mere consequence of the former holding greater relative informational value than the latter, but rather, it also bears upon the probability and predictability of the outcome with which the redundant cues are paired. The first three experiments questioned the generalisability of the redundancy effect with a partially reinforced blocked cue to the hormone change paradigm. Further investigation in Experiment 2.4 found evidence of the redundancy effect in a design where an unpredictable outcome was introduced to both the blocking procedure and the relative validity design, confirming outcome predictability as an important factor in learning about ambiguous cues. The various statistical properties of the training history proposed to differentiate learning in ambiguous situations are found to theoretically map onto the summed error learning algorithm with a competitive attention mechanism from associative analyses.

Judgment of the blocked cue as being a more likely cause of the outcome than the uncorrelated cue has been hypothesised by Uengoer et al. (2013) to reflect the higher relative informational value possessed by the blocked cue than for the uncorrelated cue. In the standard condition, the blocked cue offers exactly the same information as the accompanying blocking cue but the uncorrelated cue is less useful than the perfect predictors trained in the same compounds. Changing the reinforcement schedule for the blocking procedure from continuous to partial changes the outcome probability for both the blocking cue and the blocked cue, ensuring that they both remain equally predictive of the same outcome. Since the difference in relative informativeness is maintained between the blocked cue and the uncorrelated cue, the corresponding difference in likelihood judgment should be as well. However, unlike the persistent redundancy effect that Uengoer and colleagues (2013; Experiment 3) have observed, reducing the blocking contingencies from 100% to 50% resulted in the blocked cue being viewed as an equally or less probable cause than the uncorrelated cue in Experiments 2.1-2.3. These results suggest that absolute cue-outcome

relationship, as previously rejected by Uengoer et al. (2013), may still play a part under certain circumstances of ambiguous learning.

Relative Informativeness vs. Absolute Relationship

Contrary to Uengoer et al. (2013), the lack of redundancy effect under 50% partial reinforcement in the current experiments suggests that an explanation in terms of relative informativeness is unlikely to be simple and straightforward. One may attempt to explain the redundancy effect by appealing to not only differences in the relative informativeness between the blocked and uncorrelated cues but also to a common cue that is shared among all training and test trials. Although the predictiveness of the blocked and blocking cues are always the same, the predictiveness of the blocked cue changes relative to base rate across the different contingencies employed in the current series. In comparison, the uncorrelated cue always provides poorer utility than the reliable predictors that accompany it and usually conveys comparable predictiveness to the context or at least is closer to the predictiveness of the context than the blocked cue. The only exception where both redundant cues are equally predictive as the context is when the blocked cue is trained with a 50% contingency and the base-rate is also 50%. Therefore, fluctuations in likelihood judgment about the blocked cue may to a certain extent reflect its varying informational value relative to the context, which in turn modulate the strength of the redundancy effect.

In contrast to relative informativeness, the absolute relationship between redundant cues and the outcome offers a straightforward explanation. The higher outcome probability for the blocked cue over the uncorrelated cue gives rise to high likelihood judgment for the former and low likelihood judgment for the latter. Equating the outcome probability for the two cues makes them equally capable of activating the representation of outcome occurrence and non-occurrence, nullifying any judgement difference that would be present with imbalanced reinforcement schedules. However, taking the likelihood judgment at test to

reflect the absolute outcome probability during training is difficult to reconcile with widely replicated cue competition effects. For example, in blocking (e.g. Kamin, 1969), the 100% trained blocking cue can retard learning about the also 100% trained blocked cue, and in conditioned inhibition (e.g. Rescorla, 1969), learning about a novel cue can be hindered by the simultaneous presence of an inhibitory cue despite being followed by the outcome with 100% probability.

Outcome Predictability

Despite the focus on relative and absolute relationships as key determinants of learning, the probabilistic blocking conditions used here and by Uengoer et al. differ from the other critical contingencies in another important way. The certainty with which the outcome will or will not occur given all cues present — or outcome predictability— differs substantially for trial types involving Z compared to those involving X and Y. The blocked cue is consistently reinforced on AX trials and is thus experienced with both high outcome probability and high outcome predictability. Increased uncertainty applies to intermittent training of single trial types but may not hold true if partial reinforcement is spread across several different trial types. In the case of the uncorrelated cue, although the outcome follows the uncorrelated cue in a standard relative validity with 50% probability, it forms deterministic relationships with the outcomes and can be anticipated with perfect accuracy by referring to the specific trial type. The uncorrelated cue is thus experienced under conditions of low outcome probability but high outcome predictability. The partially reinforced blocked cue on IZ trials is not only paired with the outcome at a reduced rate but also with reduced outcome predictability. In other words, the outcome can be perfectly predicted on AX, BY, and CY trials but cannot be predicted with perfect accuracy on IZ trials.

The extent of learning may not only rely on how well the cue predicts the outcome compared to cooccurring cues (i.e. the relative informativeness), but may also rely on how

well the outcome can be predicted across training trials that contain the cue (i.e. the outcome predictability). Numerous studies from human contingency learning have revealed a positive transfer of both forms of predictive validity to the rate of subsequent learning (Le Pelley et al., 2010; Kattner, 2015). There is also evidence from the present study that increasing the relative informativeness of a cue while decreasing its outcome predictability might abolish the learning advantage expected by the relative validity effect. Notwithstanding the considerable work from animal learning experiments showing the relative validity effect—i.e. greater conditioning to the common cue from pseudo-discrimination than that from relative validity — (Wagner et al., 1968; Cole et al., 1995; Murphy et al., 2001a; Murphy et al., 2001b), rather few reports of the relative validity effect have been documented in the human learning literature. Thus, it seems reasonable to speculate that both relative informativeness and outcome predictability contribute to human contingency learning in combination. However, it should be noted that there has also been evidence that the learned predictiveness effect occurs independently of relative informativeness or outcome predictability, but relies on absolute outcome probability only (Livesey et al., 2011).

Summed Prediction Error with Attention

Computational modelling identified the summed error learning algorithm as the primary driver and attention prioritization as the secondary driver for the redundancy effect and related findings. The statistical properties of the training schedule discussed in preceding sections provide speculative explanations as to why the hallmark pattern for the redundancy effect was observed under certain circumstances but diminished in others. These statistical properties may reconcile with formal associative accounts by considering their similarities. It follows from the summed error learning rules that cues with high associative values should dominate the prediction on a given trial, and by doing so, retard learning about their companions. The summed error learning algorithm is thus sensitive to both absolute and

relative properties of the statistical relationships and a straightforward parallel cannot be drawn. The separable error learning algorithm on the other hand tracks absolute relationships, where learning is directly proportional to the probability at which an outcome follows a cue. Moreover, prioritizing attention for better predictors over poorer ones according to the Mackintosh theory or the Uengoer theory would entail comparisons of predictive power between concurrent cues. That is, allocation of attentional resources is proportionate to a cue's ability to predict the outcome relative to accompanying cues, and by how much more reliable the cue is compared to others if alpha is calculated based on associative strength. The attentional mechanisms are thus mainly sensitive to relative informativeness of cues. Results from Experiments 2.1 to 2.4 do not map cleanly onto any of the statistical properties in isolation, and it is not surprising that the models providing the best fit for the data overall employ a summed error term.

The modelling results tended to favor summed prediction error models mainly due to the fact that separable error learning rules are worse at capturing the training data. This constraint is not fully overcome with an adjustable alpha term to enhance learning about better predictors at the cost of learning about poorer predictors. That is, cooccurring cues compete less vigorously for associative strength if relying on individual error learning algorithms even with an added preferential attention mechanism in favor of good predictors. As a result, despite the seemingly parsimonious account provided by outcome probability, absolute relationships established on the basis of discrepancy between each individual cue and the outcome would at best explain the observed difference in likelihood judgment for redundant cues but would struggle to accommodate the influence of companion cues on learning.

Limitations and Future Directions

The theoretical concepts and the learning algorithms in the discussion so far capture most of the present findings, yet they still face certain issues. Most notably, they do not explain why the redundancy effect was found with 0% outcome probability in the absence of a cue but was not with 50% outcome probability on no-cue trials in Experiment 2.3. This might be because of the way in which base rate manipulation was implemented. Specifically, greater uncertainty carried with the probabilistic outcome on the supposedly context-reflecting trials might elevate overall uncertainty of the task compared to when the absence of the outcome on these no-cue trials was observed deterministically. This might increase the uncertainty around the causal status of the blocked cue, leading to an intermediate judgment tendency that reflects the ambiguity of both the blocked and the uncorrelated cues (Jones et al., 2019). Future research investigating the role of context is encouraged to take into account the level of overall uncertainty when making base rate adjustments.

While both approaches lend some credence to the assumption that all trial cues share an overlapping component, it is not without shortcomings. In fact, simulations of the traditional Rescorla-Wagner model under the common element assumption have revealed inconsistent findings. The original theorisation provided by Vogel and Wagner (2017) showed, across extended training, a diminishing yet sustained pattern of greater associative strength accrued to the blocked cue than for the uncorrelated cue. This was, however, not the case for the two papers by Uengoer and colleagues (2018, 2020). In particular, the common element model predicted an initial redundancy effect that was gradually reversed over time in the former, and failed to predict the redundancy effect with a two-phased blocking procedure in the latter. Both the transient nature of the redundancy effect with the standard design and the complete elimination of the effect with staged blocking are difficult to reconcile with the empirical observations made by these authors.

A further reason for the inability to replicate the persistent redundancy effect under partial reinforcement may lie in the different cover stories chosen by Uengoer et al. (2013) and the current study. As alluded to in previous sections, the food allergist scenario implies the assumption that the effect of allergenic foods cannot be prevented by other foods consumed together. In this respect, if a cue is indeed a valid cause, it must always be followed by the outcome regardless of the cues accompanying it. In the food allergist task, if one never observes an allergic reaction after eating foods B and Y together, it seems defensible to conclude that Y does not trigger an allergic reaction. The possibility that Y is allergenic but food B prevents its effect can be discounted (Jones et al., 2019). This decrementing effect on likelihood judgment for the uncorrelated cue is arguable not as strong in the hormone change task because it is plausible that B could take on a possible preventative role. It is conceivable that some learners in the current study entertained the belief that drug Y did not lead to the outcome on BY trials because B prevented the generative effect of Y evident on CY trials from taking place. Such an inferential explanation would suggest that the judgment difference between redundant cues depends on the formation of deductive inferences, rather than the associative processes involved. Future work is warranted to explore the possible involvement of reasoning processes by targeting deductively valid assumptions for redundant cues.

In the current experiments, the learning advantage for the blocked cue over the uncorrelated cue was demonstrably diminished under 50% partial reinforcement for the blocking contingencies. This finding lends little support to the proposal that the redundancy effect is sustained by a mere difference in relative informativeness between the critical cues (Uengoer et al., 2013). However, the characteristic judgment pattern was observed in a condition where uncertainty was equally distributed to trials involving the blocked and uncorrelated cues alongside the introduction of a probabilistic outcome to the blocking

procedure. These results were best accounted for by the summed error learning algorithm under the common context assumption (Vogel & Wager, 2017), which may be bolstered by a kind of attention preference for reliable predictors (Mackintosh, 1975; Uengeor et al., 2020). The findings highlight that neither absolute nor relative statistical properties are likely sufficient to explain the effect, but a combination of the two may be important for learning about ambiguous cues.

Chapter 3

Deductive Reasoning about Ambiguous Causes

People are sometimes likened to intuitive scientists in the way we come to understand cause and effect in our everyday lives (Kahneman & Tversky, 1982; Nisbett & Ross, 1980; Holland et al., 1986). Unlike the ideal conditions presented by a scientific experiment, which allow us to independently manipulate potential causal factors, the real world often contains highly correlated events and their causal roles in producing consequences can be hard to tease apart. When it comes to assaying the environmental cues that potentially signal meaningful events, redundancy could be said to be the norm. At a busy intersection, a change in traffic lights, sudden drop in traffic noise, cars coming to a stop, and movement of other pedestrians all provide some information signalling that it may be safe to walk across the road, even though one of these events is the cause of the others. Understanding how we learn about ambiguous, and particularly *redundant* cues, is crucial to understanding the psychological processes that allow our experiences to give rise to an understanding of cause and effect. The redundancy effect is clearly relevant to how we reason about ambiguous causes. Although much of much of this literature concerns specific predictions made by different models of associative learning, the effect—and more generally, what people learn about potential causes that are rendered ambiguous or redundant in different ways—could be highly informative about the way people form beliefs about cause and effect.

The blocking effect, in particular, has informed debate in the last few decades about how causal knowledge is acquired and how human causal learning relates to learning processes in other contexts and other species. One class of prominent explanations for human causal learning is derived from associative models traditionally developed to study Pavlovian animal conditioning, such as the influential Rescorla-Wagner model (Rescorla & Wagner, 1972). These models employ mathematical algorithms to compute changes in associative

strength on a trial-by-trial basis. It is assumed that animals, including humans, form mental representations for different events and learning progresses as connections between these representations establish and strengthen through repeated cooccurrence. An associationist account of causal learning assumes that the individual infers causation based on the extent to which the cue retrieves or activates a representation of the outcome via these connections. This idea is appealing because it provides a straightforward way of theoretically linking memory and judgment processes, using a simple inferential heuristic. However, it is also limited given that it is largely silent about the effect of the specific properties of the cue-outcome relationship, other than the statistics of their co-occurrence (e.g. De Houwer et al., 2002; Mitchell & Lovibond, 2002).

In contrast to this associative analysis, inferential explanations assign more importance to the role of controlled reasoning processes. According to what may be termed the propositionalist account of causal learning (Mitchell et al., 2009), it is thought that humans store propositional knowledge about the relationships between events, based on reasoning about observations (e.g. retrieved episodic memories), and their causal judgements reflect the strength of beliefs about propositions concerning cause and effect. In contrast to the associationist tradition, the specific properties of the relation between cue and outcome (e.g. causal mechanism, if known) are assumed to be encoded as part of the proposition. The proposition takes a form that is sensitive to the qualities of the reasoning process that has been used to derive it, which in turn depend on the individual's mental model of the processes involved, the constraining beliefs that the individual assumes to be true.

Some theorists have argued that human learning in all forms is *exclusively* propositional in nature (De Houwer, 2009; Mitchell et al., 2009). This position is contentious but irrespective of whether it is considered tenable, it is certainly possible that explicit beliefs about cause and effect are strongly determined by the inferential processes that support

propositional learning. Considerable evidence in support of an inferential approach has been documented (see De Houwer et al., 2005, for a review). Of particular relevance to this study are examples that participants' judgements are sensitive to the boundary constraints required for valid causal inferences to be made. These include ceiling effects on the observable strength of an outcome (e.g. Wu & Cheng, 1999; De Houwer et al., 2002) and the additivity pretraining effect (e.g. Lovibond et al., 2003; Livesey et al., 2019), which demonstrate that, under certain circumstances, human causal learning may be better understood by appealing to specific reasoning accounts.

Lovibond et al. (2003) noted that the ambiguity around the causal status of the blocked cue may depend on the assumptions made about effect magnitude. Consider the example of blocking. If participants assume that effects of multiple causes are additive, they can conclude from observing that the outcome on AX trials was no larger than that on A alone trials that X is not a cause. On the other hand, if participants believe that the effect is the same size regardless of the number of causes simultaneously present, then there are at least two possible causal accounts consistent with the evidence—either both cues are causes or only one of them is a cause and the other is a non-cause—which means the causal status of the blocked cue is ambiguous. Therefore, providing conditions that clearly permit magnitude additivity may remove this causal ambiguity around the blocked cue. This typically involves an explicit demonstration that the effects of single cues can combine to produce a larger effect (i.e. *additivity* pretraining). Participants are presented with two causal cues, each producing an effect of certain severity, and observe an effect of greater magnitude when these two individual causes are compounded together, such that the separate effects of individual cues sum to produce a more extreme effect. Consistent with inferential reasoning, this additivity pretraining subsequently reduces participants' judgements of the blocked cue such that the blocking effect is more pronounced (e.g. Beckers et al., 2005; Livesey & Boakes,

2004; Livesey et al., 2019; Lovibond et al., 2003). A similar result has been reported in experiments that use an explicitly *submaximal* outcome. That is, in the presence of either A or the combination of A and X the outcome occurs at the same magnitude despite the obvious possibility that it could be larger, again encouraging the participant to infer that X has no causal role in producing the outcome. Blocking under such conditions has been found to be larger than under maximal conditions, in which any further contribution of X to the outcome would not be observable (e.g. De Houwer et al., 2002; Beckers et al., 2005). These results have provided support for the propositionalist account of blocking and causal learning more generally.

Here, it should be evident that deductive reasoning plays an important role in the propositionalist account. For instance, the logic of the submaximality and additivity manipulations discussed above is equivalent to *modus tollens*; if X causes the outcome then AX should result in a larger outcome than A alone (if p then q), AX results in the same outcome as A alone (not q), therefore X does not cause the outcome (therefore not p). The formal application of deduction is permitted by allowing the individual to make certain assumptions about the evidence presented, in this case, that causes are additive and a larger outcome magnitude is observable. Whether individual learners apply formal deduction or merely some approximation of it (see Rips, 1994; Barbey & Barsalou, 2009), the process of applying inferential reasoning should increase the learner's confidence in the judgment that they make about the blocked cue because the inferential process eliminates one possible state of the world, thus reducing the causal ambiguity of the redundant cue. Indeed, when very explicit additive pretraining is provided, that appears to be the case for blocked cues. Livesey et al. (2019), for instance, found an increase in participants' rated confidence in their judgments about the blocked cue after learning the additivity rule, relative to groups given non-additivity pretraining or no pretraining at all.

Some causal learning situations may invite the use of deduction without the need for deliberate experimental manipulation, based purely on the participant's prior assumptions. In their original demonstration of the additivity pretraining effect, Lovibond and colleagues (2003) argued that this was the case for another cue competition effect to which they applied additivity pretraining. If a cue I is presented in compound with another cue J and followed by the outcome, but cue J is shown not to cause that outcome in isolation (e.g. IJ+ / J–) then cue I is normally given a relatively high causal judgement. Lovibond et al. showed that additivity pretraining did not influence the strength of this *release from overshadowing* effect. Their argument was that participants already deduce that cue I is causal based on their ability to eliminate the possibility that J is causal. The causal status of I is unaffected by outcome additivity assumptions because deductive reasoning about the causal status of the added cue is possible with or without a strictly additive outcome. It merely requires an assumption that at least one cue must be responsible for an effect. This is arguably also true of the uncorrelated cue Y in the relative validity design in Table 2.1. Since the learner observes that the outcome occurs in the presence of CY but not in the presence of BY, they may deduce that C is the only causal cue, and may do so without the need for a strictly additive outcome.

If this is indeed the case, the redundancy effect in causal learning may reflect a difference between X and Y in the ease with which the learner can deduce that the redundant cue does not cause the effect. There is some evidence that participants do not readily use deductive reasoning to judge a blocked cue even when they hold the assumptions that permit deduction. Livesey et al. (2019) found that a group given a regular blocking procedure with no pretraining produced the same blocking effect in causal judgments, and the same level of confidence in their ratings for the blocked cue, as a group given explicit *non-additivity* pretraining. Since removing the assumptions necessary for deductive reasoning had no impact on the magnitude of blocking, this result suggests the blocking effect in these

experiments was not a product of deduction in the first place. In contrast, blocking in causal judgments and confidence for the blocked cue were enhanced in an additive pretraining group despite the fact that the group given no pretraining clearly displayed additive assumptions on other tests. These results were observed using the food allergist task, in which blocking tends to be highly replicable regardless of how assumptions are manipulated prior to learning. The available evidence thus suggests that 1) deductive reasoning about the blocked cue occurs under some conditions and, when it does, it enhances blocking, but 2) participants will not necessarily engage in such reasoning without explicit forms of encouragement such as pretraining, and 3) other cognitive processes produce blocking even when deduction is not engaged (Livesey et al., 2019).

On the other hand, learning about the uncorrelated cue Y may engage deductive processes much more readily to form the belief that Y is not causal through the observation that it inconsistently leads to the outcome. The only additional constraining assumption that such deduction requires is that B is not *preventing* Y from causing an outcome that it would otherwise cause in the presence of C. If participants fail to entertain the possibility that B is preventative or decide that it is implausible, then it is relatively straightforward to conclude that the outcome observed in the presence of CY can be attributed to C and not Y (Zaksaite and Jones, 2020). Not surprisingly, inferential reasoning processes have been proposed in a number of studies as a promising candidate explanation for the redundancy effect in causal learning (e.g. Jones et al., 2019; Spicer et al., 2020). Consistent with the possibility that the uncorrelated cue Y invites deductive reasoning more readily than the blocked cue X, Jones et al. (2019) found that confidence ratings for cue Y tended to be higher than those for X.

In summary, evidence suggests that participants *can* (under the right conditions) use deduction to shape their beliefs about potential causes but do not always do so, and their propensity to use deductive reasoning may vary depending on the particular assumptions

permitted by the learning task and the manner in which the causal status of the cues has been rendered ambiguous. The aim of the current study was to test the involvement of deductive inference in the causal judgements that people make about redundant cues, focusing on blocking, relative validity and the difference between these two effects. Although the role of inferential reasoning has been examined in the blocking effect through the implementation of additivity pretraining (Lovibond et al., 2003; Livesey et al., 2019), it remains unknown whether similar manipulations of people's prior assumptions affect related causal learning phenomena such as the relative validity effect and the redundancy effect. The goal of the present research was therefore to investigate how assumptions that permit deductive reasoning contribute to these key phenomena concerning learning about redundant cues.

Experiments 3.1-3.3

Across three experiments, we presented blocking and relative validity contingencies, investigating the relative judgements made about blocked and redundant cues as well as participants' confidence for the relevant causal judgements. I retained a causal learning task in Chapter 2 in which prior causal assumptions (e.g. regarding the additivity of the outcome and the preventative properties of the cues) could easily be manipulated. Control cues were included to assess the blocking effect and the relative validity effect independently. The contingency learning design used in all three experiments is shown in Table 3.1. We explicitly manipulated participants' assumptions about magnitude additivity (Experiment 3.2) and the preventative capabilities of the cues (Experiment 3.3) to examine how encouraging and discouraging deduction may affect judgments of ambiguous redundant cues. In all other respects, the experiments were virtually identical, and thus we present the methods for the three together below.

Table 3.1

Design of Experiments 3.1, 3.2 and 3.3.

Experiment	Pretraining	Training	Likelihood ratings test	Forced choice test
3.1	(no pretraining)	Blocking:		
3.2	(additive / non-additive)	A+ AX+ DE+	X Y	X vs. Y
	M+ N+ **MN++/+**		A B C	X vs. E
	O– P– OP–	Relative Validity:	D E	Y vs. G
3.3	(preventative / non-preventative)	BY– CY+	F G H	
		FG+/– HG+/–	I J K L	
	M+ N+ MN+ O– P– OP–			
	MO–/+ NP–/+	Fillers:		
		I– J– KL–		

Note. Letters represent cues, randomly assigned to different medicine names. "+" represents hormone increase, "++" represents a large hormone increase, and "–" represents no hormone change. In Experiment 3.2, the additive group received MN++ in pretraining, while the non-additive group received MN+ in pretraining. In Experiment 3.3, the non-preventative group received MO+ and NP+ in pretraining, while the preventative group received MO– and NP– in pretraining.

Most demonstrations of the redundancy effect in causal learning have used the food allergist task. Zaksaite and Jones (2020) developed the hormone change paradigm to investigate the redundancy effect and demonstrated its utility as an alternative. This is particularly important because, as Jones et al. (2019) noted, participants' prior experience with food allergies might make preventative relationships seem particularly implausible;

foods rarely, if ever, prevent the allergies caused by other foods. If deduction plays a role in the causal inferences made about the uncorrelated cue Y then the use of the food allergist task may well have facilitated many of the previous demonstrations of the redundancy effect in human causal learning (Uengoer et al., 2013; Uengoer et al., 2019; Uengoer et al., 2020; Zaksaite & Jones, 2017; Jones & Zaksaite, 2018; Jones at al., 2019). Here, we required a task in which we could manipulate participants' assumptions about how drug effects may combine or prevent one another, in a clear and relatively straightforward way. Therefore, we used the same simplified version of the task as in Chapter 2, in which hormone levels always either remain unchanged or increase as a consequence of drug administration.

This hormone change task is relatively new and has been used in relatively few studies (Zaksaite & Jones, 2020; Spicer et al., 2020) to demonstrate the redundancy effect. These studies made the critical comparison between the blocked cue X and the uncorrelated cue Y, but did not include the control cues necessary to independently test for blocking of the X cue and relative validity for the Y cue. In the one experiment in Chapter 2 that included these control cues (Experiment 2.2), we found no evidence of either blocking or relative validity. Moreover, while prediction certainty was proposed to dissociate from causal judgment of redundant cues, none of these studies explicitly measured confidence for X and Y. Experiment 3.1 therefore sought to confirm the validity of the hormone change paradigm in the context of human predictive learning and specifically, the redundancy effect, the blocking effect and the relative validity effect when using binary outcomes (i.e. no change or increase in hormone levels). Following Jones et al. (2019) and Livesey et al. (2019), likelihood estimates and self-reported confidence were both measured. Confidence judgments were of particular interest in this study because deductive reasoning accounts make clear predictions about how confidence should diverge from causal judgements. Measuring both likelihood and confidence judgments allows potential correlations between estimated likelihood that a

cue is a cause and certainty with which this estimation is made to be revealed and determine whether the redundancy effect is due to a lack of certainty for X. Additionally, as has been argued in several papers (Perales & Shanks, 2008; Jones et al., 2019; Barberia et al., 2021), midrange likelihood ratings may not solely reflect the perceived causal power for a particular cue, but rather, it may be influenced by the certainty with which such judgments are produced. In this view, intermediate ratings equally well reflect the impression of an ambiguously judged strong cause and the impression of a confidently judged moderate cause. As a complementary measure, we also included forced choice tests for likelihood judgment and confidence in which participants were asked to choose between two critical cues (e.g. between X and Y).

We expect the blocked cue X to be regarded as a less probable cause than the overshadowed control E because it adds no new information about the outcome to that already supplied by the other cue A in the same compound while E has the same predictive power as D from the DE compound. The overshadowed cue will always have a higher conditional probability of causing the outcome than the blocked cue and thus it could be considered rational for participants to show a blocking effect no matter what cognitive process they use to generate their judgments (Livesey et al., 2013). However, the decision process that they use should affect their relative confidence in their judgments of X and E. If participants deduce that X cannot be causal and thus eliminate this possibility then they should be more confident about their judgment of X than about E. Equal confidence in the two judgments may be an indication that the same decision process has been applied to both (even if their likelihood judgements are different). For instance, if they were to apply the same conditional probabilistic reasoning process to both cues, or if they were to rely on an intuitive sense of associative retrieval fluency for both cues then they might be equally confident about X and E while still demonstrating a blocking effect in their likelihood ratings.

Alternatively, participants may conflate likelihood ratings and confidence in a way that results in the same information affecting both ratings. If that were the case then we would expect to see the same differences between cues expressed on both measures.

With regard to the relative validity effect, cue Y which is less correlated with the outcome than co-occurring cues, should be judged as a less likely cause of the outcome than G, which is equally correlated with the outcome as its associates F and H. Differences in confidence for judgements about cues Y and G may be affected by the type of uncertainty associated with each. Whereas there is ambiguity about Y because it is never seen in isolation, the outcome associated with G cannot be accurately predicted, suggesting that at least some of the uncertainty is unresolvable. An observant learner might therefore recognise that the judgement is made under aleatory uncertainty (i.e. risk) and not just ambiguity in their own knowledge (e.g. see Kozyreva & Hertwig, 2021, for further discussion). In addition, however, confidence in Y may be affected by whether participants deduce that it cannot be a cause of the outcome. Participants who fail to entertain, or who explicitly *discount*, the possibility that cues can prevent the occurrence of the outcome could deduce from BY– trials that Y is not causing the outcome. In this case, confidence would be high for Y. Choice of learning task may play an important role in whether participants hold the right assumptions for this deduction to be permitted. If the hormone change task did indeed make participants entertain the possibility of preventative relationships, then the status of Y would be rendered ambiguous by the consideration of preventative B.

The redundancy effect is characterised by higher causal ratings (or expectation of the outcome) for the blocked cue X than for the uncorrelated cue Y. It was anticipated that a pattern of results consistent with this direction would be observed alongside the blocking effect and the relative validity effects. On the basis of confidence ratings reported in Jones et al. (2019), we would expect that higher confidence for Y than for X would be observed.

However, this may be dependent on the extent to which participants apply deductive reasoning to cue Y and it was assumed that our use of the hormone cover story may be less likely to encourage deductive reasoning if participants did indeed consider preventative as well as generative relationships.

Method

Participants

All three experiments used first- and second-year psychology students from the University of Sydney participated in exchange for partial course credit. Following similar causal learning work in our lab (Don & Livesey, 2017; Livesey et al., 2019), participants were excluded if their prediction accuracy in the last half of any training phase was below 60%.

For Experiment 3.1, in which participants' assumptions were not manipulated with pretraining, tested 70 participants. One participant was removed for failing to meet the training accuracy criterion, leaving N=69 for further analysis (50 females, mean age=20.69, *SD*=6.27). This gave us over 90% power to detect a redundancy effect of the magnitude reported by Zaiksate & Jones (2020, Experiment 2; dz=.43) using a very similar causal learning task.

For Experiments 3.2 and 3.3, which employ between-subjects pretraining manipulations, we aimed to test 100 participants each. In Experiment 3.2, 101 participants completed the experiment. Two were removed for failing to pass the 60% learning threshold in the last half of training, leaving 50 (38 females, mean age=19.68, *SD*=2.28) in the additive group and 49 (35 females, mean age=19.88, *SD*=2.80) in the non-additive group in the final data analysis.

In Experiment 3.3, 98 participants were recruited. Four participants failed to pass the 60% learning threshold in the last half of main training and were excluded from analyses,

leaving 48 (36 females, mean age=19.85, *SD*=4.60) in the non-preventative group and 46 (34 females, mean age=20.46, *SD*=2.90) in the preventative group.

Design

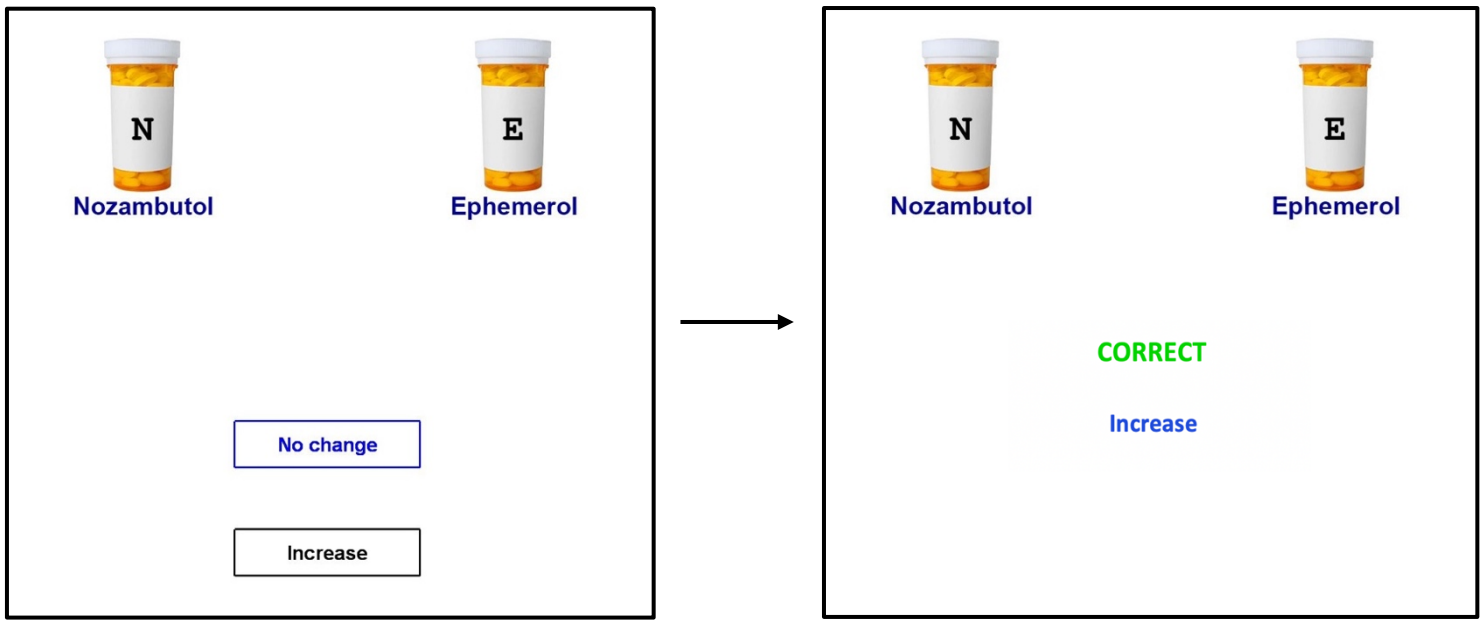
The design of all three experiments is shown in Table 3.1, and was identical across the experiments with the exception of the between-subjects pretraining manipulations used in Experiments 3.2 and 3.3. For the training phase, in addition to the standard preparation for the blocking effect and the relative validity effect, filler trials were added to balance the frequency of trials that resulted in outcome present and absent. This change was made because cue competition effects such as the blocking effect and the redundancy effect may be susceptible to the probability of the outcome occurring on any given trial (see Jones et al., 2019; Livesey et al., 2013, as well as in Chapter 2 of this thesis).

In keeping with Chapter 2, two types of tests were administered, the primary likelihood ratings test and the complementary forced choice test, measuring both perceived probability of a cue being a cause and confidence associated with likelihood judgements. Although there was no intention to apply associative learning models to the results of this chapter (the primary function of the choice test in Chapter 2 was for model fitting), I retained the test as a complementary exploratory measure.

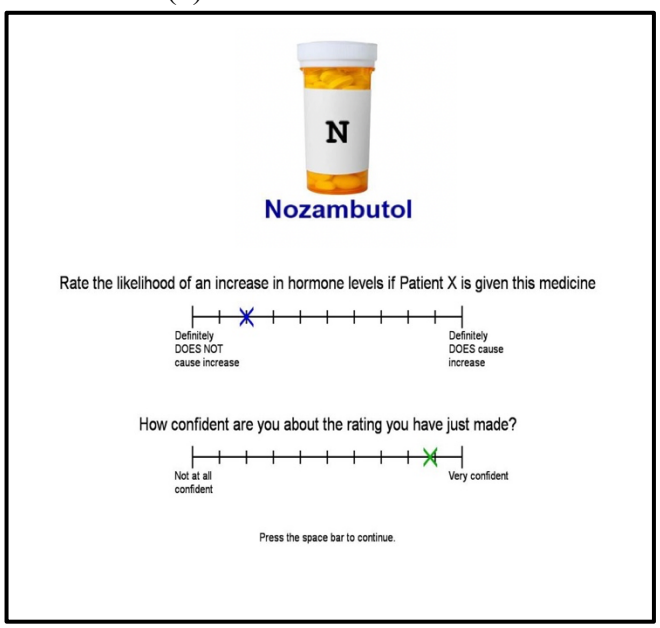
Figure 3.1

Example screenshots of (a) the prediction and feedback phase during training, (b) the ratings test, and (c) the forced choice test in Experiment 3.1.

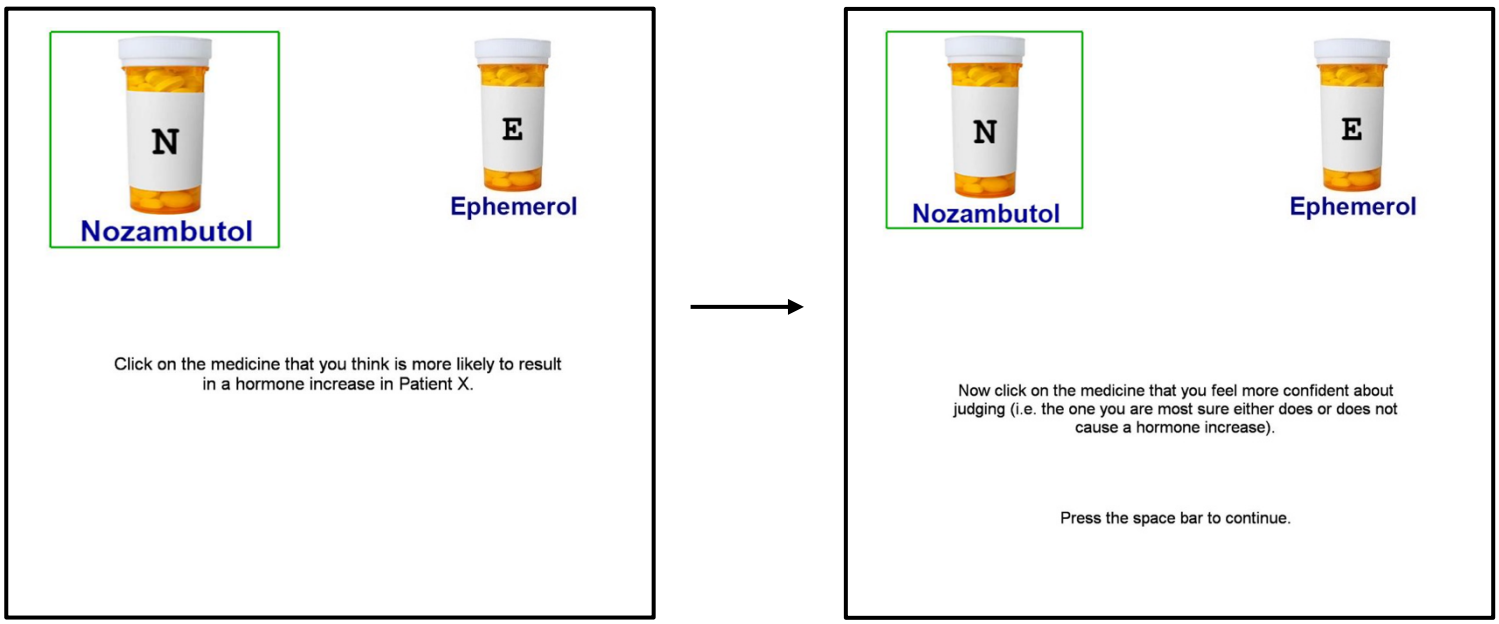
(a)


N
Nozambutol
E
Ephemerol
No change
Increase
N
Nozambutol
E
Ephemerol
CORRECT
Increase

(b)


N
Nozambutol
Rate the likelihood of an increase in hormone levels if Patient X is given this medicine
Definitely DOES NOT cause increase
Definitely DOES cause increase
How confident are you about the rating you have just made?
Not at all confident
Very confident
Press the space bar to continue.

(c)


N
Nozambutol
E
Ephemerol
Click on the medicine that you think is more likely to result in a hormone increase in Patient X.
N
Nozambutol
E
Ephemerol
Now click on the medicine that you feel more confident about judging (i.e. the one you are most sure either does or does not cause a hormone increase).
Press the space bar to continue.

Apparatus and Stimuli

The experiment was programmed using the PsychToolbox extension (Kleiner et al., 2007) for MATLAB. The experiment was completed on desktop computers situated along two walls of laboratory rooms. The experimental cues were fictitious medicine names each beginning with a unique letter from A to R. These included *Aspetur, Broncin, Chrurin, Dioxnyl, Ephemerol, Felicium, Gambutrol, Hyronalin, Impbatine, Jamitol, Krayoxx, Lithorol, Metazine, Nozambutol, Ontapelium, Plycidox, Quelinum and Rodvuccial*. Medicines were randomly allocated to serve as cues in Table 3.1. The layout of the task is illustrated in Figure 3.1. The outcome options were 'no change' and 'increase' for all training trials. Each cue was presented as a label on a 300 x 300 pixel image of a medicine bottle on the upper half the screen. On trials where there was a single cue, the cue was presented in the middle. On trials with two cues, one was located on the left and the other was located on the right (counterbalanced across trials within each block of training). The possible outcomes were presented as choices for the participant to predict and were presented in boxes with one above the other appearing on the bottom half the screen. After each outcome selection, corrective feedback was displayed at the centre of the screen signified by the word 'correct' if accurate or 'incorrect' if inaccurate with the correct outcome 'no change' or 'increase' appearing below. Participants in the Additive pretraining condition of Experiment 3.2 also observed a 'large increase' outcome on MN++ pretraining trials (note that their prediction was considered correct if they predicted 'increase' on these trials, as only the 'no change' and 'increase' prediction options were ever provided to participants). The words 'correct' and 'incorrect' were presented in green and red, respectively. The corrected outcome 'no change' was presented in black and the corrected outcome 'increase' was in blue. The task occurred against a white background. Participants proceeded through instructions by pressing the space bar and responded using the mouse.

Procedure

The task was set in a research context where participants assumed the role of a medical researcher attempting to determine whether there would be an increase or no change in hormone levels in Patient X after the administration of various medicines. All participants except those in the non-preventative condition in Experiment 3.3 were instructed to consider the preventative nature of some medicines. Specifically, they were told that in situations where two medicines lead to no hormone change, two potential interpretations are possible: (1) one medicine may lead to an increase and the other may prevent this increase from occurring, thus cancelling out their effects, or (2) neither medicine leads to an increase.

The three experiments differed only in the administration of the pretraining phase (i.e. before learning the blocking and relative validity contingencies). In Experiment 3.1, there was no pretraining phase. In Experiment 3.2, we implemented outcome additivity versus non-additivity pretraining. Before the start of pretraining, participants were given detailed instructions to direct their attention to the severity of the outcome. For the additive group, participants were told that if two medicines that produce a mild hormone increase when taken alone, they will produce a larger hormone increase when taken together. For the non-additive group, participants received the instruction that these two medicines will cause the same mild hormone increase when taken together as they each did when consumed separately. After being given these extra instructions, participants observed these changes in hormone levels in the pretraining trials.

In Experiment 3.3, we implemented preventative versus non-preventative pretraining. The instructions and procedure used the non-additive condition of Experiment 3.2 participants, with the addition that participants were taught different rules regarding preventative relationships between cues and the outcome. Those in the non-preventative group were instructed and shown that cues were incapable of preventing the effect of a valid

cause. In pretraining, when a cue associated with the outcome (e.g. M or N) appeared with a cue that was associated with the absence of the outcome (O or P), the two cues in compound still resulted in a hormone increase of the same size as that would take place if the causal cue was taken on its own. On the other hand, those in the preventative group were instructed that some cues were capable of such prevention, and were instead shown that these cue compounds resulted in no change.

On each trial in the pretraining and training phases, either a single or a compound of two medicines was presented and participants predicted whether they would lead to hormone increase by clicking either of the binary outcomes. The spatial location of cues was counterbalanced and trial orders were randomised within each block. In Experiments 3.2 and 3.3, participants completed 4 blocks of the pretraining phase, each comprising 2 presentations of each trial type. After pretraining, participants proceeded directly to the main training phase without break or further instruction. During the training phase, participants completed 8 blocks of trials, each comprising 2 presentations of each trial type.

The two test phases then followed. On each trial of the ratings test, two rating scales were provided, asking participants to rate the likelihood that individual medicines would lead to a hormone increase on one scale, and then their confidence in their ability to judge the likelihood that the medicine caused a hormone increase on the other scale. Participants rated the likelihood that specific medicines would cause an increase in hormone levels on a scale ranging from “Definitely DOES NOT cause increase” to “Definitely DOES cause increase”, with scores ranging from 0 to 100. The participant was then asked “How confident are you about the rating you have just made?” and rated their confidence on a scale from “Not at all confident” to “Very confident”, with scores ranging from 0 to 100.

After the ratings test, participants were asked to make a series of two-alternative forced choices about pairs of cues. On each trial in this cue choice test, two cues were presented and

participants were asked to first choose which medicine they thought was most likely to result in a hormone increase in Patient X, and then indicate which medicine they felt more confident about judging. Each pair of medicines was presented twice, counterbalancing the left/right spatial location of the cues.

Data Analysis

Training accuracy calculated for inclusion criterion only involved trials which consistently predicted the presence or the absence of the outcome. Those with the outcome occurring on 50% of the occasions were excluded as neither outcome could be deemed correct for these partially reinforced trials. The main learning phenomena of interest were the redundancy effect, the blocking effect and the relative validity effect. The analyses reported concern key comparisons between the relevant target cues for these phenomena. All training and test data are available in supplementary materials. All three effects were assessed through ratings tests and cue choice tests via pair-wise comparisons (E vs. X; G vs. Y; X vs. Y) in three separate t-tests. All tests were two-tailed and significance was taken to be .05.

Results

Pre-training and Training

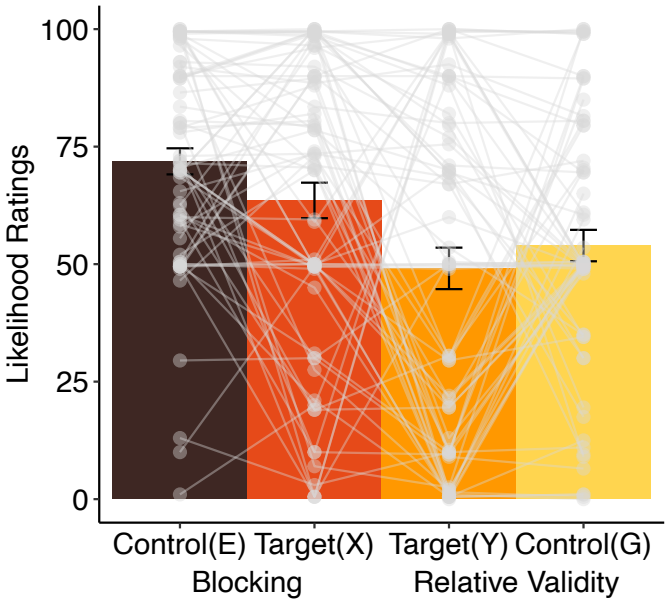
Cue-outcome contingencies were learned rapidly across training blocks in all three experiments, with participants reaching an accuracy above 90% in the last half of pretraining (Experiment 3.2 Additive = 98.17%, Experiment 3.2 Non-additive = 97.87%, Experiment 3.3, Preventative = 93.95%, Experiment 3.3 Non-preventative = 96.10%) and the last half of training (Experiment 3.1 = 94.52%, Experiment 3.2 Additive = 93.16%, Experiment 3.2 Non-additive = 95.09%, Experiment 3.3, Preventative = 93.54%, Experiment 3.3 Non-preventative = 93.78%). For further details regarding performance in the pretraining and training phases, please refer to supplementary materials.

Ratings Test

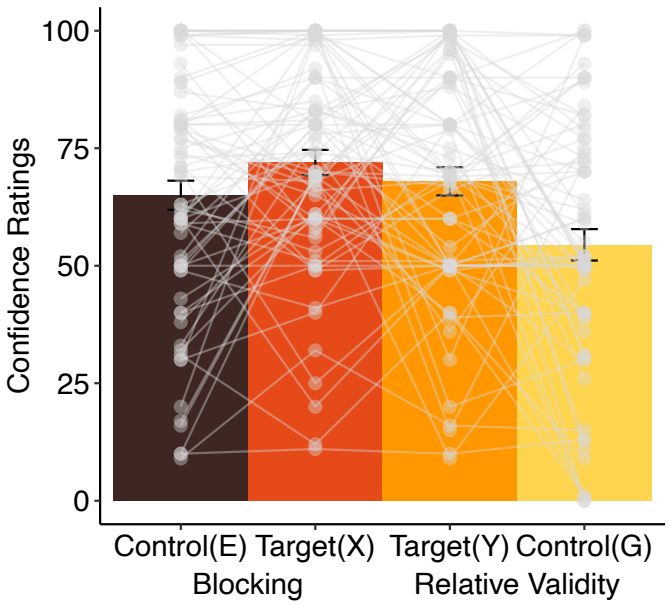
Figure 3.2

(a) Mean likelihood ratings and (b) Mean confidence ratings on the ratings test of Experiment 3.1. (c) Mean likelihood ratings and (d) Mean confidence ratings on the ratings test of Experiment 3.2. (e) Mean likelihood ratings and (f) Mean confidence ratings on the ratings test of Experiment 3.3. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

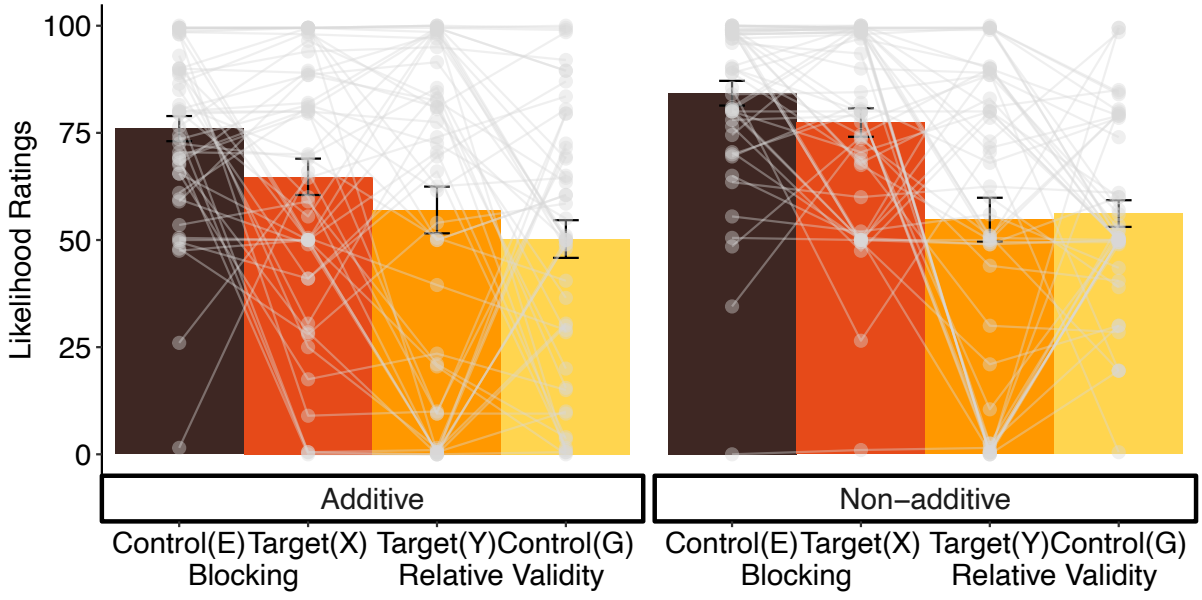
(a) Experiment 3.1 Likelihood Rating



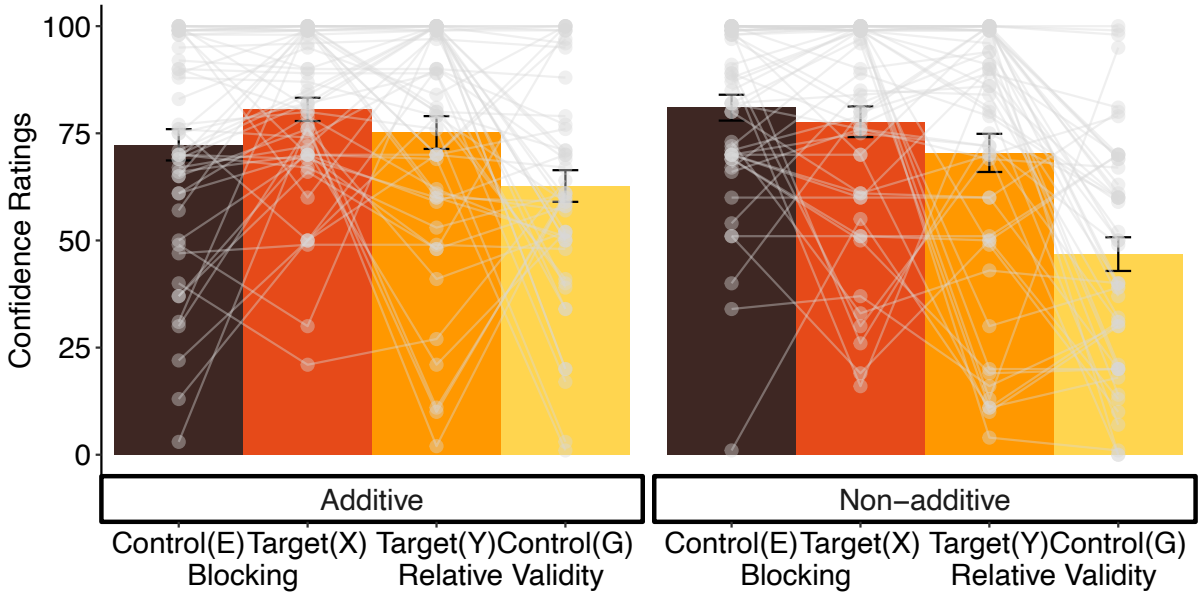
(b) Experiment 3.1 Confidence Rating



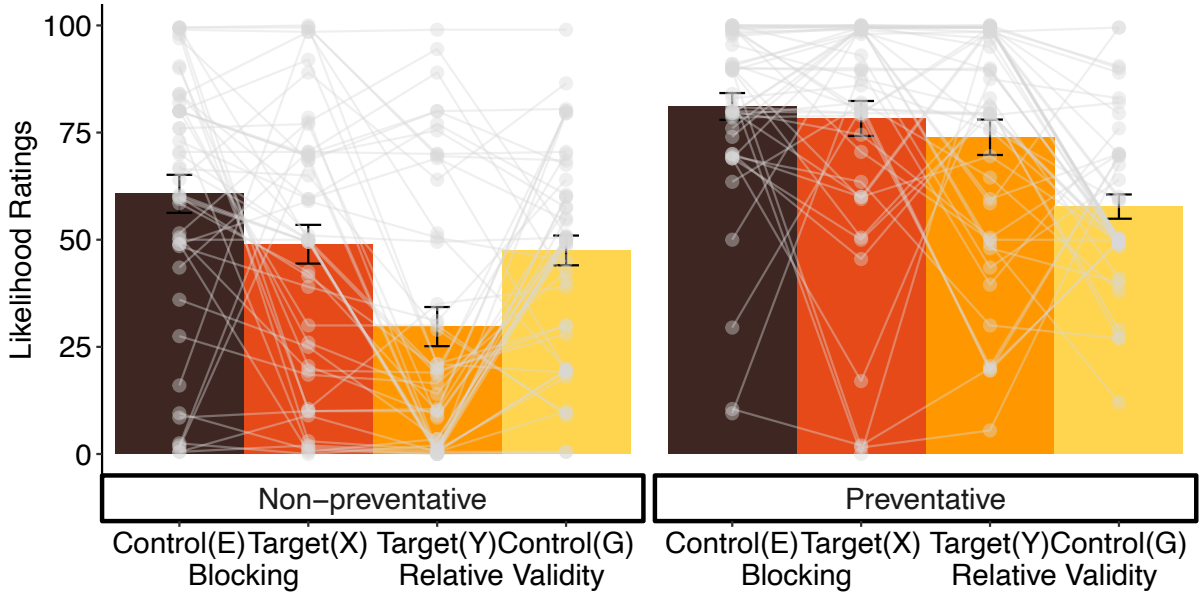
(c) Experiment 3.2 Likelihood Rating



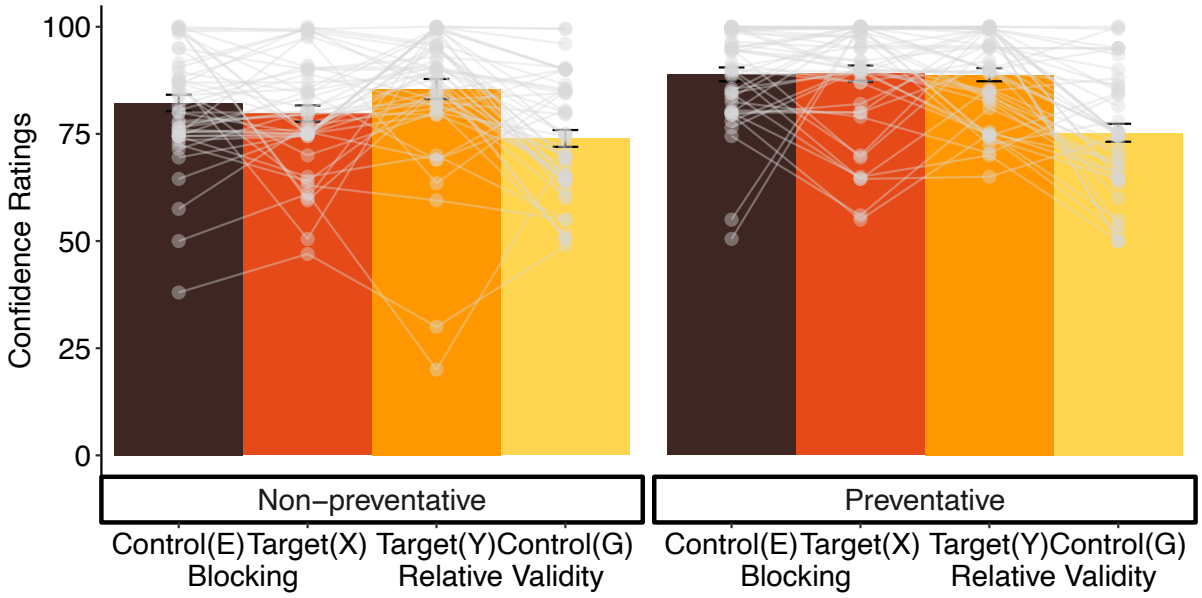
(d) Experiment 3.2 Confidence Rating



(e) Experiment 3.3 Likelihood Rating



(f) Experiment 3.3 Confidence Rating



Mean likelihood ratings and mean confidence ratings for critical cues in all three experiments are shown in Figure 3.2. Please see supplementary materials for complete figures including all cues for Experiments 3.1-3.3.

For Experiment 3.1, to determine whether differences in mean likelihood ratings matched the three effects of interest, pair-wise comparisons revealed significantly higher ratings for E than for X, $t(68)=2.24$, $p=.028$, $d=.270$, indicating a blocking effect, but no difference in ratings was found comparing Y to G, $t(68)=1.03$, $p=.309$, $d=.123$ (i.e. no relative validity effect). Ratings were significantly higher for X than for Y, $t(68)=2.84$, $p=.006$, $d=.342$, indicating a redundancy effect. These same three comparisons can be found for each experimental group across Experiments 3.1-3.3, along with the equivalent comparison in confidence ratings, in Table 3.2. The lack of rating difference between Y and G raised the possibility that participants did not learn to distinguish between B and C as expected. However, comparison between these two cues revealed higher ratings for C (M=68.72, SD=26.33) than for B (M=27.55, SD=26.46), $t(68)=8.37$, $p<.001$, $d=1.007$. Comparisons within pairs of cues demonstrated that participants were more confident about X than E, and more confident about Y than G but equally confident about X and Y (see Table 3.2).

For Experiment 3.2, pairwise comparisons were made between critical cues across the additive group and the non-addictive group using three separate 2x(2) mixed measures ANOVA. Results showed that X was on average given significantly lower ratings than E, $F(1,97)=11.42$, $p=.001$, $\eta_p^2=.105$, indicating the presence of an overall blocking effect. Interestingly, the average ratings for X and E were higher in the non-additive group than they were in the additive group, $F(1,97)=6.87$, $p=.010$, $\eta_p^2=.066$. Independent samples t-test showed that ratings were significantly higher in the non-additive group than in the additive group for X $t(97)=2.34$, $p=.021$, $d=.471$, and E, $t(97)=2.01$, $p=.047$, $d=.404$. It is noteworthy

that the interaction between cue type and group did not reach significance, $F(1,97)=.677$, $p=.413$, $\eta_p^2=.007$. Both groups exhibited a significant blocking effect (see Table 3.2). However, inconsistent with the relative validity effect, likelihood ratings were not lower for Y than for G, $F(1,97)=.44$, $p=.511$, $\eta_p^2=.004$, nor was the cue type x group interaction significant, $F(1,97)=1.02$, $p=.314$, $\eta_p^2=.010$. Moreover, X was rated as a more likely cause than Y averaged over group, $F(1,97)=17.10$, $p<.001$, $\eta_p^2=.150$, indicating the presence of an overall redundancy effect. This higher likelihood ratings for X over Y was significantly more pronounced in the non-additive group than in the additive group, $F(1,97)=4.12$, $p=.045$, $\eta_p^2=.041$. Only the non-additive group exhibited a significant redundancy effect (see Table 3.2).

The same analyses performed on confidence ratings revealed that the higher confidence ratings to X than to E were more marked in the additive group than in the non-additive group, $F(1,97)=6.28$, $p=.014$, $\eta_p^2=.061$. Indeed X only received significantly higher confidence ratings than E in the additive group (see Table 3.2). Participants were on average more confident in their likelihood judgements made for X than those made for Y, $F(1,97)=4.16$, $p=.044$, $\eta_p^2=.041$, and were more confident about Y than about G, $F(1,97)=26.22$, $p<.001$, $\eta_p^2=.213$. All other effects were non-significant.

Likelihood ratings in Experiment 3.3 were analysed in the same way as for Experiment 3.2. Results showed that X was on average given higher likelihood ratings than Y, $F(1,92)=8.47$, $p<.001$, $\eta_p^2=.084$, and the higher ratings to X than to Y was significantly more marked following non-preventative pretraining than following preventative pretraining, $F(1,92)=5.45$, $p=.022$, $\eta_p^2=.056$, and indeed X was rated significantly higher than Y only in the non-preventative group (see Table 3.2). Regarding the blocking effect, ratings were significantly higher for E than for X, $F(1,92)=4.52$, $p=.036$, $\eta_p^2=.047$, and higher in the

preventative group than in the non-preventative group, $F(1,92)=28.30$, $p<.001$, $\eta_p^2=.235$. The cue type x group interaction was however non-significant, $F(1,92)=1.71$, $p=.194$, $\eta_p^2=.018$, despite their only being a significant blocking effect in the non-preventative group (see Table 3.2). Regarding the relative validity effect, ratings for Y and G were on average higher in the preventative group than in the non-preventative condition, $F(1,92)=41.56$, $p<.001$, $\eta_p^2=.311$. There was a significant cue type x group interaction, $F(1,92)=25.35$, $p<.001$, $\eta_p^2=.216$, driven by a significantly higher ratings for G than for Y in the non-preventative group *and* significantly higher ratings for Y than for G in the preventative group (Table 3.2).

The same analyses were performed on confidence rating for Experiment 3.3. Comparing X versus Y, results revealed that Y was on average judged with more confidence than X, $F(1,92)=4.23$, $p=.043$, $\eta_p^2=.044$, and participants in the preventative group were more certain about X and Y on average than those in the non-preventative group, $F(1,92)=7.36$, $p=.008$, $\eta_p^2=.074$. Importantly, there was a significant cue by pretraining group interaction, $F(1,92)=4.99$, $p=.028$, $\eta_p^2=.051$, driven by significantly higher confidence about Y than X in the non-preventative group but not the preventative group (Table 3.2). For the X versus E comparison, participants felt more confident judging the likelihood of X and E on average following non-preventative pretraining than following preventative pretraining, $F(1,92)=13.64$, $p<.001$, $\eta_p^2=.129$. The effect of cue type, $F(1,92)=.71$, $p=.401$, $\eta_p^2=.008$, and the cue type x group interaction, $F(1,92)=.94$, $p=.335$, $\eta_p^2=.010$, were however non-significant. For the comparison between Y and G, Y was on average given higher confidence ratings than G, $F(1,92)=66.49$, $p<.001$, $\eta_p^2=.420$, but the main effect of group was not significant, $F(1,92)=.32$, $p=.574$, $\eta_p^2=.003$, nor was the interaction effect significant, $F(1,92)=.17$, $p=.684$, $\eta_p^2=.002$.

Table 3.2

Cue competition effects and associated differences in confidence for individual groups

Comparison	Group (df)	Likelihood ratings				Confidence ratings			
		t	*p*	*d*	BF_{10}	*t*	*p*	*d*	BF_{10}
Blocking	E1	2.24	.028	.270	1.364	2.35	.022	.283	1.704
E – X	E2 Additive	2.51	.016	.355	2.584	2.65	.011	.374	3.497
	E2 Non-additive	2.37	.022	.339	1.964	0.97	.338	.138	.241
	E3 Non-preventative	2.22	.031	.320	1.450	1.20	.234	.174	.309
	E3 Preventative	0.66	.516	.097	.196	-.086	.932	-.013	.161
Relative	E1	1.03	.309	.123	.219	-3.32	.001	-.399	18.138
validity	E2 Additive	-1.134	.262	-.160	.282	-2.45	.018	-.347	2.295
G – Y	E2 Non-additive	.261	.795	.037	.160	-4.85	<.001	-.693	1461.466
	E3 Non-preventative	3.73	<.001	.539	52.786	-4.56	<.001	-.658	582.796
	E3 Preventative	-3.39	.001	-.500	20.886	-6.11	<.001	-.901	70976.752
Redundancy	E1	2.84	.006	.342	5.181	1.24	.219	.149	.275
effect	E2 Additive	1.74	.088	.246	.626	1.41	.166	.199	.387
X – Y	E2 Non-additive	3.85	<.001	.550	74.131	1.48	.145	.212	.432
	E3 Non-preventative	4.68	<.001	.676	849.334	-2.37	.022	-.343	1.971
	E3 Preventative	.90	.372	.133	.234	0.14	.892	.020	.161

Forced Choice Test

Mean proportions of choices on the forced choice test for all three experiments, are shown in Figure 3.3. Choice of the most likely cue to cause a hormone change are illustrated in the bars labelled ‘Cause’ and choices of the cue about which the participant is most certain are illustrated in the bars labelled “Conf”. Choice proportions for the first cue from the XE pair, the YG pair and the XY pair were analysed through three separate non-parametric one-sided sign tests, with proportions below 50% coded as negative, 50% chance level proportion coded as 0 and proportions above 50% coded as positive.

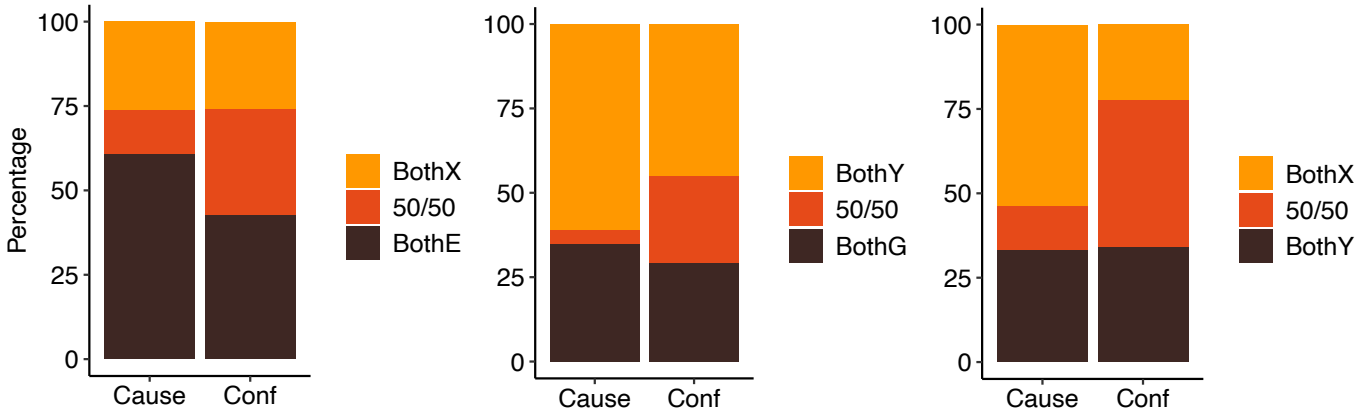
For Experiment 3.1, the results revealed that E was selected significantly more often as the more likely cause than X, p=.001, indicating the presence of a blocking effect. However, Y had a higher probability of being chosen as the more likely cause over G, p=.012, showing the reverse of the relative validity effect. X was chosen as the more likely cause than Y, p=.046, favouring the redundancy effect. Participants were more certain about

Y than G, $p<.001$, but were similarly confident about X and Y, and X and E. It should be noted that participants were numerically more confident about E than about X from inspecting Figure 3.3a, which seems to oppose the higher confidence for X than for E revealed on the ratings test.

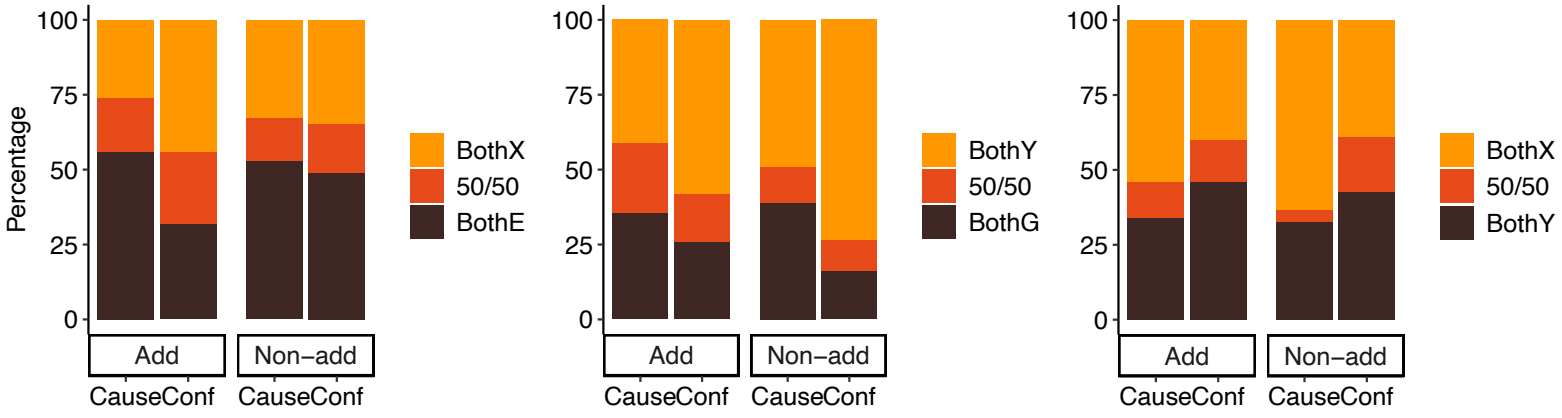
Figure 3.3

Mean percentage for choosing the same cue once and twice from the XE pair, the YG pair and the XY pair on the forced choice test for causal judgement and confidence in (a) Experiment 3.1, (b) Experiment 3.2, and (c) Experiment 3.3. Higher percentage of 'Cause' indicates higher likelihood that a given cue is chosen as a cause once or twice and higher percentage of 'Conf' indicates higher certainty in the ability to judge the effects of a given cue.

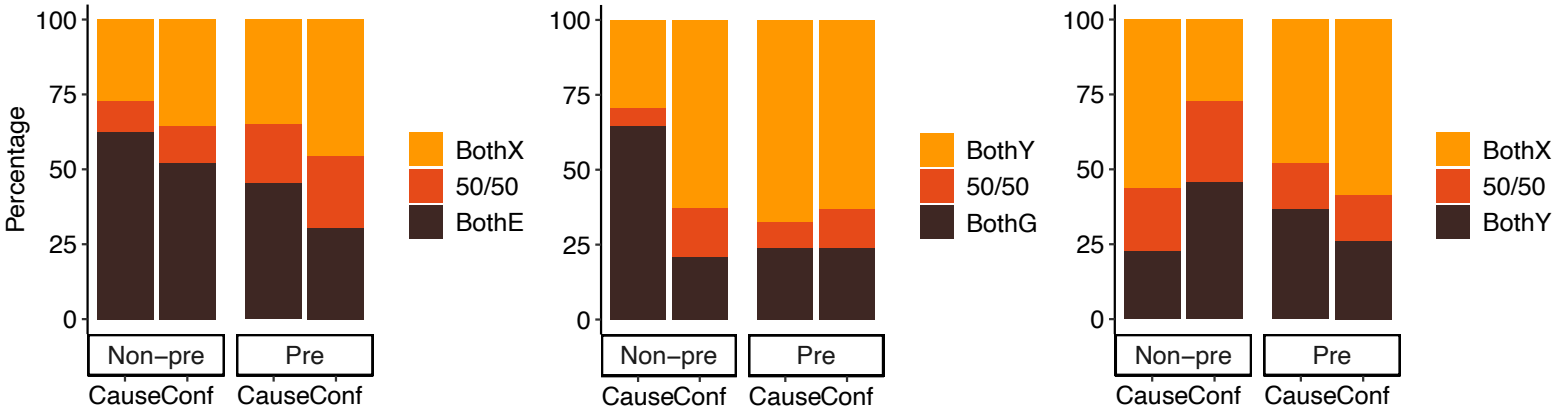
(a) Experiment 3.1



(b) Experiment 3.2



(c) Experiment 3.3



For Experiment 3.2, choice proportions for the first cue from the XE pair, the YG pair and the XY pair were analysed through two sets of non-parametric one-sided sign tests for cause and confidence, respectively. Results revealed that X was chosen significantly less often compared to E as the more probable cause in the additive group, p=.014, but not in the non-additive group, p=.082. The relative validity effect in the form of greater choice probability for G over Y was significant in neither group, with ps>.7. Inspecting Figure 3.3b, it appears that a greater proportion of participants chose X over E as the more certain cue in the additive group than in the non-additive group, however, neither the additive group, p=.209, nor the non-additive group, p=.895, showed significantly higher confidence for X than for E. The sign tests for self-reported confidence revealed that participants chose Y as the cue that they felt more confident judging over G in both the additive group, p=.010, and the non-additive group, p<.001. Moreover, X had a higher probability of being selected as the more likely cause than Y in the non-additive group, p=.020, but not in the additive group, p=.087. Confidence for X and Y was not significantly different in either group, with ps>.3.

The same analyses were applied to the results of Experiment 3.3, shown in Figure 3.3c. Results showed that X was less often chosen as the valid cause than E in the non-preventative group, p=.007, but not in the preventative group, p=.256, demonstrating evidence for a

significant blocking effect in the former but not the latter group. Choice proportion for Y was significantly lower than G in the non-preventative group, p=.008, but was significantly greater than G in the preventative group, p=.001, suggesting that the relative validity effect was significant in the former group but reversed in the latter. Moreover, X was more often selected as the more likely cause than Y in the non-preventative group, p=.007, but not in the preventative group, p=.261, indicating the presence of the redundancy effect in the former but not the latter group. Analyses of certainty choices showed that Y was more frequently chosen as the more confident cue than G in both the non-preventative group, p=.001, and the preventative group, p=.003. Y was more often selected as the less confident cue than X in the preventative group, p=.012, but not in the non-preventative group, p=.088, suggesting that certainty about Y dropped after entertaining a preventative assumption. Although inspecting Figure 3.3c suggests that participants were slightly more confident about E than about X in the non-preventative group, this higher confidence did not reach significance, p=.140.

Discussion

Across three experiments, we investigated how people learn about two types of causally ambiguous cue (the blocked cue X from a blocking design and the uncorrelated cue Y from a relative validity design), and specifically how they judged the likelihood of each cue causing an effect and their confidence in being able to judge that likelihood. These experiments made use of a causal scenario that made it relatively easy for us to manipulate assumptions that should be critical for reasoning about these cues. In particular, we manipulated the assumptions that should be necessary to confidently eliminate the possibility that the cue was a cause of the outcome, based on deductive inference. For each of the cue competition effects of interest in this study (blocking, relative validity, and the redundancy effect), we will first discuss how people performed in this task when no manipulations of

these assumptions were made (Experiment 3.1), before considering how the pretraining in Experiments 3.2 and 3.3 affected these judgements.

Blocking

In Experiment 3.1, blocking was evident in both likelihood ratings, where the overshadowed E received higher ratings than the blocked X, as well as in choice proportions, where E was more frequently chosen than X as a more likely cause of the hormone increase. In terms of self-reported confidence, that is, the degree to which participants could be certain that their likelihood judgments were correct, participants did not give similar confidence ratings to X and E as participants did in Jones et al. (2019) but were more certain about X than about E. The greater certainty with respect to the causal property of the blocked cue raises the possibility that deductive reasoning for X might have been used by a certain proportion of participants, leading to an elevated confidence for the non-causality of X and a robust blocking effect. According to a deductive inferential explanation for blocking, it is possible to conclude the non-causal nature of X from the fact that X presented together with a known cause A results in the same effect as A alone. When comparing the confidence level for X with E, the status of E is less certain as it can be either D or E that causes the outcome. Livesey et al. (2019) argued that while participants may hold the prior assumption that the effect of multiple causes should add up to an effect of larger magnitude, they do not tend to actively engage in deductive reasoning without being deliberately encouraged. However, the evidence on which that conclusion was based came from a different causal context to this one (i.e. the food allergist task) and it remains to be seen whether their results generalise broadly. In the absence of further investigation, it remains unclear whether deductive inferential processes were involved and, if so, why such reasoning processes were facilitated in this experiment.

Experiment 3.2 explored whether explicitly providing or removing conditions that permit deductive reasoning about the blocked cue X, through outcome additivity versus non-additivity pretraining, would influence the way redundant cues are judged as possible causes. The blocked cue X was rated as being a less likely cause of the outcome than the overshadowed cue E by both additive and non-additive participants. It should be noted, however, that ratings to both X and E were lower following additive pretraining compared to non-additive pretraining, pointing to an effect of additivity pretraining on both cues, not previously reported in other studies. These likelihood ratings for X were made with higher confidence than E following additive pretraining whereas X and E were rated with the same confidence in the non-additive group, suggesting that additive assumptions may help disambiguate the causal role of X while non-additive assumptions do not. The forced choice test revealed that the effect of blocking was present following additive pretraining but was reduced after non-additive pretraining. The results are partially consistent with past literature (Lovibond et al., 2003; Beckers et al., 2005; Livesey et al, 2019) that demonstrated the additivity pretraining effect on blocking. Rather than completely abolishing blocking, these studies found a numerically weaker yet statistically significant blocking effect after removing magnitude additivity assumptions. Importantly, the lower causal or likelihood ratings for X in the previous studies were given with higher confidence after additive than after non-additive pretraining. While the present experiment did not find precisely the same pattern, X was judged with higher confidence relative to E with additive pretraining compared to non-additive pretraining.

The pretraining conditions used in Experiment 3.3 were designed to manipulate an assumption critical to reasoning about cue Y and thus we had no *a priori* expectations about their effect on blocking. Somewhat surprisingly, ratings for X and E were higher in the preventative group. Although the interaction between cue and group was not significant, the

preventative group was the only group across the three experiments that did not exhibit a significant blocking effect. In both groups, deductive reasoning about X should have been restricted by implementing non-additive pretraining. However, preventative pretraining may have unintentionally made a different form of deduction possible for X by encouraging participants to assume that the two constituent cues of a compound must both be valid causes (i.e. if X did not *prevent* the outcome of A on AX trials, then it must be causal).

Relative Validity

Measured as higher likelihood ratings for the common cue G from a pseudo-discrimination (FG+/-, HG+/–) than its counterpart Y from a relative validity (BY–, CY+), the relative cue validity effect was observed on neither the causal ratings test nor the forced choice test in Experiment 3.1. Although ratings were slightly lower for Y than for G, this weak bias was not statistically reliable. More surprisingly, Y was selected as having a greater probability of leading to the hormone level increase than G from the Y vs. G forced choice test, indicating the opposite of the relative validity effect. The pattern of results observed in both groups in Experiment 3.2 were similar to Experiment 3.1, although this time differences between Y and G were not significant on either test.

While this result at first glance appears incompatible with the existing body of research on stimulus relative validity (Wasserman, 1990; Baetu et al., 2005; Callejas-Aguilera & Rosas, 2010), it may be reconciled by taking into account the methodological differences. The hormone change paradigm invites the participant to consider preventative relationships between cues and the outcome (the instructions provided at the beginning of the task explicitly encourage it). These changes suggest that Y might have been seen as a cue whose increasing effect on hormone levels observed on CY trials was hindered by the accompanying B on BY trials. An important consequence of considering preventative relationships is that it would increase the uncertainty with which the causal status of Y can be

determined. If prevention was considered impossible, then Y could not be a cause as it did not lead to the outcome on BY trials. If prevention was considered possible, then it could either be the case that Y was causal but its effect was prevented by B, or Y was not causal. The causal status of G, on the other hand, would be less affected by preventative assumptions and should remain ambiguous as the relationships between FG and GH compounds and the outcome were inconsistent.

In both Experiments 3.1 and 3.2, self-reported confidence revealed that participants were more confident about Y than they were about G. A plausible explanation for this result is possible in terms of the nature of cue-outcome relationship, where deterministic relationships are judged with more confidence than probabilistic relationships which continue to convey some uncertainty even when they are well learned (e.g. Tannenbaun et al., 2017; Kozyreva & Hertwig, 2021). Presented with two cues possessing similar causal ambiguity, Y and G, participants may have more confidence making likelihood judgment for Y whose outcome can be predicted with certainty than for G whose outcome can never be accurately predicted.

Experiment 3.3 directly manipulated an assumption that should permit or prevent participants from confidently eliminating cue Y as a cause. Non-preventative pretraining allows the participant to deduce that Y is not a cause of hormone change when they observe BY– trials, whereas this deduction is not valid after explicitly preventative pretraining. After non-preventative pretraining, we found a relative validity effect (significant lower likelihood ratings for Y than for G) for the first time in this study, compared to a significant effect in the opposite direction after preventative pretraining.

Redundancy Effect

In line with existing literature, in Experiment 3.1, ratings to X were substantially higher than those to Y on the likelihood ratings test, suggesting that participants regarded the

blocked cue from a blocking treatment as a more likely cause of the hormone increase than the uncorrelated cue from a relative validity. This numerical difference was present in all experiments, though it was not significant in the Additive pretraining group in Experiment 3.2 or the Preventative pretraining group in Experiment 3.3. The presence of the redundancy effect was further evidenced by the higher probability of choosing X as the more likely cause in the X vs. Y choice test. Again, this numerical difference was generally present across experiments but not significant after Additive pretraining (Experiment 3.2) or Preventative pretraining (Experiment 3.3).

In contrast, results for confidence were somewhat unexpected. Jones and associates (2019) found that participants were less confident about X than about Y and proposed that the redundancy effect may reflect this difference in certainty resulting from the ambiguous causal status of X. However, Experiment 3.1 did not observe such differential confidence for X and Y despite the presence of a reliable redundancy effect. This result thus provides little support for their proposal. The only condition to produce reliably stronger confidence ratings for Y than for X was the non-preventative pretraining group in Experiment 3.3, in which we actively encouraged participants to hold assumptions that make it possible to eliminate cue Y as a cause of the hormone increase.

It is noteworthy that the two conditions that did not show systematic evidence of the redundancy effect are the conditions in which we tried to 1) enhance deductive reasoning about X (Additive, Experiment 3.2) and 2) reduce deductive reasoning about Y (Preventative, Experiment 3.3). This finding is consistent with the hypothesis that the redundancy effect is elicited in situations where it is easier to deduce that Y is non-causal than it is to deduce the non-causal nature of X. In Experiment 3.2, making the non-causal deduction easier to perform for X by encouraging additivity assumptions (i.e. X did not add anything useful to A) substantially weakened the effect that would otherwise be obtained in the absence of any

pretraining. A similar outcome was achieved in Experiment 3.3 when we made the non-causal deduction more difficult to perform for cue Y, by strongly encouraging preventative assumptions.

Individual variation in reasoning and judgment: K-means clustering analysis

Previous work using causal learning procedures has shown that not all participants reason in the same way despite observing the same contingency information and reading the same contextual constraints on cause and effect (e.g. Don et al., 2020; Shanks & Darby, 1998). In the context of deductive reasoning in causal learning, it is also clear that individuals do not always engage in deduction just because they possess the necessary assumptions to do so (Livesey, et al., 2019). If causal judgment of ambiguous cues is influenced by deductive inferences then it raises the question of how consistently individuals apply such inferences; are these effects carried by a subset of individuals and, if so, how do manipulations of the reasoning constraints change the probability of an individual using deduction?

The inferential theory of learning delineates a clear inverse relationship between confidence and likelihood judgment for redundant cues under circumstances that permit the individual to deduce that the cue is not causal. The theory thus identifies a style of behaviour as a likely indication of deductive reasoning taking place. That is, confidence should be elevated as deduction eliminates the possibility that the ambiguous cue may be a potential cause. Henceforth, we refer to this pattern of behaviour as *confident elimination* and label individuals who demonstrate this behaviour pattern as *confident eliminators*.

In line with the inferential explanation, if the observed difference between the blocked cue and the uncorrelated cue is attributable to reasoning processes being easier to apply for one cue than the other, then we should expect to find greater evidence of confident elimination for Y than for X, as well as changes in the extent of this pattern depending on the assumptions encouraged (and discouraged) by pretraining. At a group level, the analyses

reported so far did not reveal a consistently greater certainty for Y over X under situations where X was regarded as a more likely cause than Y, as one would expect if differences in confident elimination were producing the redundancy effect. In fact, the negative relationship between confidence and likelihood ratings was only observed in the non-preventative group in Experiment 3.3. However, this pattern should not be taken at face value without further analysis at the individual level.

As deduction produces a discrete rather than a continuous shift in behavioural pattern, the group-level null hypothesis significance test that assumes one homogeneous effect may be misleading when the underlying population contains qualitatively different relationships between likelihood judgment and perceived confidence, particularly if only a subset of individuals use deduction. It might still be the case that there are more confident eliminators of cue Y than cue X, and that the proportions of confident eliminators changes according to the constraints established using instructions and pretraining, but that these patterns are obscured by participants who do not engage in deduction. If the diminished redundancy effect in the additive condition compared to the non-additive condition, and in the preventative condition compared to the non-preventative condition is reflective of the differing ease with which deductive reasoning can be used, these effects should be reflected in the proportions of individuals who are classified as confident eliminators for each cue.

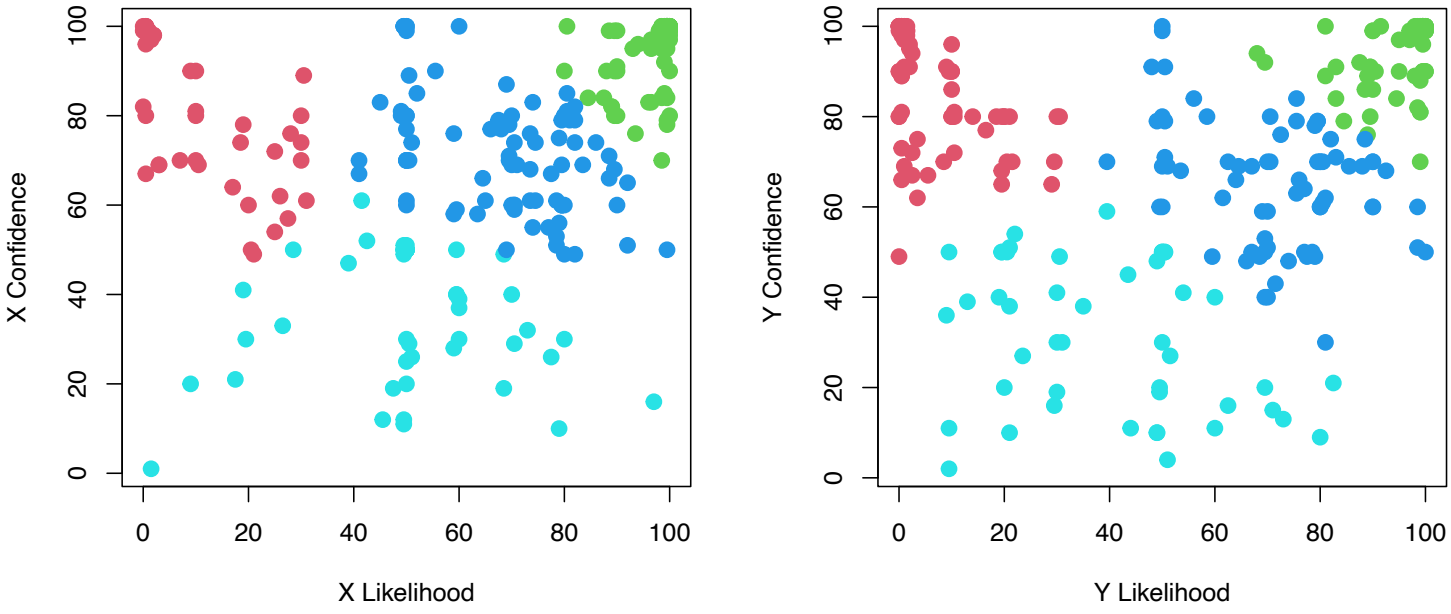
If the assumptions held by participants influenced the probability of them engaging in deduction in a manner described by the inferential account, then the group of participants who made non-causal deductions should display a pattern of judgment that is distinctively different from the rest of the cohort who did not. Based on this assumption, we used a K-means clustering analysis, which categorises respondents into discrete clusters, on data collapsed across all three experiments. This clustering algorithm should identify a group of confident eliminators who gave high confidence ratings to their low likelihood judgment for

each redundant cue. If confident elimination were indeed the result of deductive reasoning, then the magnitude of the redundancy effect should reflect a difference in the number of confident eliminators belonging to X and Y clusters. According to the inferential explanation, there should be more confident eliminators in the Y cluster compared to the X cluster in conditions that promoted deduction about Y, equal numbers in both clusters in conditions that promoted deduction for both X and Y (or neither X nor Y), and more in the X cluster compared to the Y cluster in conditions that promoted deduction about X.

Figure 3.4

K-means clustering with 4 centres plotting likelihood ratings against confidence ratings for (a) X and (b) Y.

(a) (b)



Likelihood ratings were plotted against confidence ratings for the critical cues X and Y as illustrated in Figure 3.4. The optimal number of clusters was selected according to the elbow criterion (Syakur et al., 2018) which showed a drastic decrease in sum of square error at K=4 for both cues. Therefore, data points were categorised into four groups. Confident

eliminators who deduced the non-causal nature of redundant cues in line with an inferential account is of most theoretical interest. In this regard, individuals in the top-left cluster (coloured red) were analysed further as they gave low likelihood ratings with high certainty, showing judgments consistent with the pattern hypothesised under causal deduction. To account for K-means randomness, the classification that produced the lowest total within-cluster sum of squares from 100 repetitions of the analysis was chosen as the optimal clusters. The frequency and percentage of confident eliminators in each condition are summarised in Table 3.3.

Table 3.3

Distribution of participants in the X cluster, Y cluster, in both X and Y clusters, and in neither X nor Y cluster across the groups of Experiments 3.1-3.3. Numbers outside the brackets indicate the frequency of confident eliminators and numbers inside the brackets indicate the percentage of total participants in each group.

Group	X cluster & Y cluster	X cluster only	Y cluster only	Neither	Total N
E1	4 (5.80%)	7 (10.14%)	18 (26.09%)	40 (57.97%)	69
E2 Additive	5 (10.00%)	2 (5.00%)	7 (14.00%)	36 (72%)	50
E2 Non-additive	1 (2.04%)	0 (0.00%)	11 (22.45%)	37 (75.51%)	49
E3 Non-preventative	15 (31.25%)	1 (2.08%)	15 (31.25%)	17 (35.42%)	48
E3 Preventative	2 (4.35%)	2 (4.35%)	4 (8.70%)	38 (82.61%)	46
All Experiments	27	12	55	168	262

The distribution of confident eliminators across conditions was examined with the chi-square test for equality of proportions, which revealed that the proportion differed depending on the experimental manipulation for both the X cluster, $X^2(4, N=262)=20.76$, $p<.001$, and the Y cluster of interest $X^2(4, N=262)=31.17$, $p<.001$. Subsequent between-group comparisons showed that, consistent with the manipulation implemented in Experiment 3.3, there were more confident eliminators for Y in the non-preventative group relative to the preventative group, $X^2(1, N=94)=22.27$, $p<.001$. However, the number of confident eliminators for X following additive pretraining was only numerically greater than that following non-additive pretraining in Experiment 3.2, $X^2(1, N=99)=3.29$, $p=.070$. It is worth noting that preventative pretraining also decreased the number of confident eliminators for X, $X^2(1, N=94)=7.11$, $p=.008$.

To examine the influence of confident elimination on the redundancy effect, paired samples t-tests were conducted for participants in both X and Y clusters, and in neither cluster. In these two clusters, the blocked and uncorrelated cues have been subjected to the same k-means clustering. Results revealed that the redundancy effect was significant for those falling outside both clusters, $t(167)=2.79$, $p=.006$, $d=.240$, and was in the right direction but did not reach significance for those at the intersection of the two clusters, $t(26)=.98$, $p=.337$, $d=.170$.

If confident elimination was a decision-making strategy that some individuals are generally more inclined to use than others across all ambiguous causal situations, then among the subset of participants who employed it, one may expect the X and the Y clusters of interest to be of comparable size with a large overlap between the two. Chi-square test of independence indicated that confident eliminators for X and Y are not independent, $X^2(1, N=262)=31.17$, $p<.001$. Calculation of phi coefficient indicated a positive correlation between the two classifications, $\phi=.342$. These results suggest that confident elimination for

X and Y are related, that is, there is a greater-than-chance tendency for participants to be classified as confident eliminators for both or for neither cue. However, confident elimination was more prevalent for Y than for X. On the basis of this analysis, it must also be assumed that alternative processes are usually invoked in the judgment of the cues since the majority of participants did not appear to use confident elimination (this is particularly clear for cue X).

If confident elimination is assumed to be a behavioural pattern generated as a result of engaging in deductive reasoning for the redundant cues, then these findings suggest that the redundancy effect is driven by participants' stronger tendency to use deduction for Y than for X. Importantly, the overlap between the two clusters represents a situation where the tendency to draw deductive inference for X and Y is equalised to some degree. Under such a situation, the redundancy effect was reduced. This result corroborates the hypothesis that the lower likelihood judgment for Y than for X reflects the greater ease with which reasoning processes can be engaged for Y compared to X. However, the observation of a significant redundancy effect among individuals who never engaged in confident elimination points to alternative explanations for the judgment difference between X and Y.

It should be reiterated that analyses performed here were restricted to the cluster for which the inferential account has made clear a priori predictions about their judgment pattern (i.e. confident elimination). The other three clusters appeared to have similar properties across the analyses of X and Y. That is, in both analyses there was 1) a cluster that gave very high likelihood *and* confidence ratings, 2) a cluster that gave moderately high confidence ratings and likelihood ratings in the upper half of the scale (ranging from 50 to 100), and 3) a cluster that gave low confidence ratings and widely varying likelihood ratings. It is quite possible that individuals in these clusters were organised together in other theoretically

meaningful ways. However, since we did not make *a priori* predictions about these behavioural profiles, they are not the focus of discussion here.

General discussion

The current series of experiments explored the role of inferential reasoning processes in the redundancy effect and the judgments of ambiguous cues from which it is derived. It has been proposed that the higher likelihood judgement made about the blocked cue X than about the uncorrelated cue Y reflects the differing levels of certainty associated with judgments for these cues (Jones et al., 2019). The current research lends support to this proposal by further arguing that the readiness to apply deductive inference substantially influences the magnitude of the redundancy effect. Three experiments were conducted to evaluate this claim. Experiment 3.1 provided a conceptual replication of the redundancy effect in the hormone change paradigm, with both likelihood and confidence ratings at test. Experiment 3.2 created a situation where valid deductive inferences were explicitly valid for X, established through additive pretraining, or explicitly invalid for X, through non-additive pretraining. The redundancy effect was found to be present among non-additive participants but negligible among additive participants. Experiment 3.3 compared a situation where deductive reasoning was encouraged for Y through prior learning of non-preventative assumptions with a situation where deductive reasoning was discouraged for Y through pretraining of preventative relationships. The redundancy effect was evident among non-preventative participants but was abolished among preventative participants.

Taken together, these findings suggest that the redundancy effect is underpinned, at least partly, by inferential processes that serve to disambiguate the causal status of X and Y. Where deductive reasoning is permitted, participants are able to deduce the non-causal nature of the redundant cues by entertaining additive assumptions for X and non-preventative assumptions for Y and are thus more confident in their judgments; where necessary premises

to draw causal inferences are removed, the ambiguity around the causal status of the redundant cues leads to participants judging them as moderate causes and reporting low confidence for their judgments. Importantly, the critical cues in the redundancy effect are associated with different levels of difficulty with which deductive reasoning can be employed. Deduction appears to have been made particularly easy for the uncorrelated cue Y by the default assumption that a causal cue must consistently lead to its paired outcome and the tendency to neglect preventative influences from other accompanying cues. This may be particularly true in the commonly used food allergist task. But even in paradigms that facilitate preventative assumptions, such as the hormone change task, some participants seem to persist in their belief that Y must be non-causal because the outcome was absent on BY trials.

The pervasiveness of the simple deduction for Y is contrasted with the difficulty to draw deductive inferences for X. It appears to be rather uncommon to deduce that X is a non-causal cue even when the participant's assumptions allow such deduction to take place (Livesey et al., 2019). That is, while additivity assumptions might be entertained by default, participants do not necessarily deduce X to be non-causal. Learning of non-additive rules may have discouraged learners from engaging in reasoning processes to infer the noncausality of X. However, this explicit discouragement was not a necessary condition to observe the higher likelihood judgment for X than for Y, as demonstrated in Experiment 3.1 and in other reports of the redundancy effect in the literature where no explicit discouragement was given.

The present research therefore suggests that it is partly the differing levels of readiness with which deductive reasoning can be engaged for X and Y that leads to Y being judged as a less probable cause than X. In this respect, the pretraining manipulations in Experiments 3.2 and 3.3 may be viewed as artificially creating circumstances under which the readiness to

make deductions about X and Y was roughly equalized (additive group and preventative group) versus circumstances under which the readiness differed (non-additive group and non-preventative group). Additive pretraining in Experiment 3.2 encouraged deduction for X so that confident elimination could be applied to both X and Y. Preventative pretraining in Experiment 3.3 discouraged deduction for Y so that confident elimination was unlikely for either X or Y. In both cases, the redundancy effect was no longer evident in participants' causal judgments.

Although, in Experiment 3.2, confidence did not differ between X and Y with additive or non-additive pretraining, participants did feel more certain about the causal status of X relative to E when deduction was encouraged than when deduction was discouraged. Similarly, prior learning of non-preventative relationships in Experiment 3.3 reinforced the idea that cues are incapable of preventing the outcome of another cue. As a result, participants became even more certain about Y than about X and deemed X as a more likely cause than Y. On the other hand, preventative pretraining obfuscates the causal status of Y, with the result being that X and Y were treated as equally ambiguous (in terms of likelihood *and* confidence ratings).

One striking result from the current series of experiments is the lack of the relative validity effect. Y was seen as a less probable cause than G only after explicit non-preventative pretraining which removed entirely the possibility that B may prevent the effect of the otherwise generative Y. One reason for this lack of relative validity lies in the choice of cover story. Almost all previous attempts (Baetu et al., 2005; Callejas-Aguilera & Rosas, 2010; Wasserman, 1990) to demonstrate the relative validity effect were made with the food allergy task, which, by the inherent nature of the scenario, excludes any preventative influences from other cues. The present study, on the other hand, adopted a task that allows consideration of preventative relationships. Furthermore, it is noteworthy that among all

studies concerned with the redundancy effect, only one study prior to my work (Uengoer et al., 2013) included the full preparation for both the blocking effect and the relative validity effect, and one study included the full preparation for the blocking effect (Jones et al., 2019), and again, these two studies used the food allergist task. This casts doubt on whether the blocking and relative validity effects would have been found if proper controls were included in other prior studies. Therefore, while the exact reason for the absence of the relative validity effect is unclear, it is not at odds with other redundancy effect studies in the literature.

It is worth-noting that the pseudo-discrimination control in the present set of experiments offered an opportunity (not provided in most other redundancy effect studies) for the participant to observe probabilistic cue-outcome relationships. With all contingencies being deterministic in nature, participants in previous studies may have assumed that all cues either always signal the presence or the absence of the outcome. Learning that some causes lead to their effects consistently while others intermittently lead to the outcome may encourage the consideration of the possible causal nature of Y. However, results indicated that participants continued to regard Y as an unlikely cause (i.e. less likely than X) in the absence of explicit preventative pretraining discouraging deduction. Future studies are encouraged to examine the relative validity effect more extensively with other causal learning tasks (i.e. other than the food allergist task).

The current study explored the underlying inferential processes based on a pattern of data observed in judgment and confidence. A limitation with this approach is that different types of cognitive processes resulting in the same observable pattern may be involved. It is thus unclear whether one or several processes are responsible for the patterns of causal judgments. Indeed, the data suggest that deductive inference plays an important role in determining the relative judgment of X and Y even though the majority of participants show no sign of actually using this process when we looked for judgments consistent with

confident elimination. While there is compelling evidence that deduction influences likelihood judgment by clarifying causal ambiguity around a redundant cue, it is less clear what processes are engaged by learners in uncertain situations that do not facilitate deduction.

One possibility is that the strength of associative memory may form the ground on which causal judgment is based (Thorwart & Livesey, 2016). According to associative learning models, it is assumed that experienced events are stored in an associative network as mental representations and relationships between them as links connecting concurrent events. Without any encouragement for deduction, the associative retrieval of the outcome by the presented cue may serve as the primary source of evidence that learners rely on to estimate the causal status of that cue. The fact that X is always followed by the presence of the outcome means that the presentation of X will bring to mind the representation of the associatively connected outcome to some degree (though perhaps modestly so because of the competition generated by cue A), leading to X being seen as a moderate cause. Judgment for Y may also be susceptible to associative memory retrieval, but because learners are disposed to deductive reasoning for Y, associative processes only exert a subtle influence. Note that Y activates the mental representation of both the outcome and its absence, thus some accounts based on associative memory also predict the same low likelihood judgment for Y as that arrived at through deduction (see Uengoer et al., 2013). However, the significantly improved likelihood judgment for Y by preventative assumptions in Experiment 3 suggests that deductive reasoning plays a major role in determining the causal status of Y (see supplementary materials).

Similar ideas have been expressed in decision-making literature as the fluency of explicit memory retrieval. Memory is viewed by these theorists as a strategic tool that taps indirectly into the strength of cue-outcome association via encapsulated frequency information, enabling fast inferences with little cognitive effort (Hertwig et al., 2008). One

way to test the idea that X more strongly encodes and activates the mental representation of the outcome than Y does would be to conduct a memory test at the end of the experiment (e.g. Greenaway & Livesey, 2020). Furthermore, some authors have attributed differences to the differential attention based on the learning history of the X and Y cues (Jones & Zaksaite, 2018; Uengoer et al., 2019), which remains a potential reason why the cues differ in the way they are judged.

Instead of ascribing the redundancy effect to the variation in one continuous psychological variable, the findings reported in this study suggest that judgments made in one situation may be computed by multiple cognitive processes. This approach is distinctly different from the one taken by statistical or associative explanations that assume a unitary psychological process. Although manipulating the limiting assumptions to make deductive inference valid appears to have only encouraged a minority of participants to engage in deductive reasoning, a substantial influence on judgments of ambiguous cues was observed and the redundancy effect can be essentially removed by making deductive reasoning clearly valid for both cues. Even though it is clear that other cognitive processes must be involved in forming judgements about cause and effect, the differences between X and Y in this case might be a result of people applying deduction more consistently for Y than for X. Future work is warranted to explore the types of conditions that lead to this type of inferential reasoning, the types of individuals that are more inclined to use this type of inferential reasoning, and the qualities of causal learning about ambiguous cues when these processes are not engaged.

The redundancy effect represents a particular instantiation of learning under uncertainty which directly compares two ambiguous cues in the same design. In line with an inferential account of causal learning, assumptions about magnitude additivity and preventative relationships modulated likelihood judgment about the blocked cue and the

uncorrelated cue by either permitting or prohibiting deductive reasoning. Supported by self-reported confidence associated with likelihood judgment, deduction appears to remove the causal ambiguity around the uncertain cues, leading to definitive judgment of non-causality. It is argued that under ambiguous circumstances, likelihood judgment is derived from inferential processes such as deductive reasoning where appropriate, and may be based on alternative evidence such as the strength of associative memory where such processes are not engaged.

Chapter 4

Protecting Theory about Redundant Cues

Observing coincidences between events, like a predictive cue and a subsequent outcome, usually leads us to update our knowledge about the relationship between those events. There are multiple complex factors that control the rate at which this updating occurs. One of those factors, which has been the central topic in a wealth of research, is the presence of other competing cues. In human causal learning paradigms, multiple cues are usually present concomitantly to compete for an association with the outcome. These cues vary in the predictive power they possess with respect to the outcome, and consequently, enter into learning to various degrees based on their usefulness. In the redundancy effect, while X tends to be judged as a more likely cause than Y, it has recently been revealed in one study that confidence in these judgments appears to dissociate from perceived causal likelihood with a greater degree of uncertainty attached to the predictions evoked by X than those by Y (Jones et al., 2019). This presents an interesting situation in which learners regard X as a stronger predictor of the outcome but they are more certain about their judgment of Y. How then would learners update their knowledge given new information that bears on the relationship between these cues and the outcome given that X and Y both differ in causal likelihood and causal ambiguity?

The redundancy effect has sparked renewed interest in the mechanisms underlying differential learning about redundant cues trained in separate cue competition paradigms, however, rather little is known about the fate of redundant cues when presented together for further training. Given that X and Y differ in both causal judgment and confidence judgment, it is unclear whether combined training of XY will continue to exhibit a learning bias on the basis of these factors. Assuming that contingency judgment in the form of causal likelihood reflects the strength of association that X and Y accrued during blocking training and relative

validity training, respectively, the difference in associative strength may contribute differentially to subsequent learning of the XY compound. On the other hand, if future learning hinges more on the causal certainty with which the cue-outcome associations were acquired, the varying levels of confidence may also produce asymmetrical learning between X and Y in the XY compound. It is therefore the overarching objective of the current chapter to assess the impact of previous training history of the redundant cues on their subsequent learning in the same compound.

Rescorla (2001) evaluated the relative contribution of constituent cues with different starting associative strengths to subsequent learning of a compound. To do so, a compound conditioning procedure was devised. In Stage 1, A and C were trained as excitors while B and D were neutral. In Stage 2, AB together were paired with either outcome presence (AB+) or outcome absence (AB−). At test, outcome expectation elicited by AD and BC were compared against each other. Any difference in learning between the two test compounds was thought to be generated by the unequal associative change that A and B underwent during compound training. Rescorla found that reinforcing an excitatory cue and a neutral cue together resulted in a larger increase in associative strength for the cue that was farther away from the asymptotic associative strength, that is, the neutral B underwent greater associative increment than the excitatory A during excitatory compound training. Conversely, inhibitory training of the same AB compound led to a larger decrement in associative strength for the excitatory A than for the neutral B. The major conclusion from this study was that cues holding different initial associative values undergo uneven associative changes.

These results are difficult to reconcile with associative learning models governed by a common prediction error but may be accommodated by theories that use individual error learning rules. Prediction error, the discrepancy between anticipation and observation, is widely considered as a crucial determining factor in associative learning and has become the

foundation on which an influential class of learning models are predicated. Broadly, error-driven theories can be categorised into common error term models where the amount of associative change is regulated by the degree to which the received outcome is different from the aggregated expectation made based on all elements of a compound, and individual error term models where learning is dependent on the extent of mismatch between the prediction engendered by each constituent cue and the actual outcome. The critical difference thus lies in the ability of these models to account for variations in the distribution of associative change across elements of a compound. Common error term models assign an equal amount of associative change to all members of the same compound regardless of differences in initial associative state and are therefore unable to explain the unequal share in associative change as observed by Rescorla (2000, 2001) and later studies (e.g. Bradfield & McNally, 2008; Leung & Westbrook, 2010). Individual error term models, on the other hand, allow changes in associative strength for different components to be mediated by their individually calculated prediction error. For example, a neutral cue or an inhibitor will gain a larger increment than an excitor on reinforced trials because the outcome is more discrepant with the prediction elicited by the neutral cue or the inhibitor than that elicited by the excitor. Conversely, a neutral cue or an inhibitor will suffer from a smaller decrement than an excitor on non-reinforced trials because they predict the absence of the outcome more accurately than the excitor. These results suggest that cues trained in the same compound undergo changes in associative strength that are directly proportional to their uniquely elicited prediction error. In light of these findings, Rescorla proposed a hybrid model based on the product of common and individual prediction error. The common prediction error was retained because of its widespread usefulness in explaining other independently observed learning phenomenon such as blocking (e.g. Kamin, 1969) and conditioned inhibition (e.g. Rescorla 1969).

As an alternative, Rescorla (2001) speculated about the possibility of retaining the common error term with appropriate elaboration on the function translating associative change into performance. He argued that the influence of a common error that applies to all cues in a compound may be modulated by the extent to which each individual cue deviates from the asymptotic associative strength supportable by the trial outcome. With this additional assumption in place, cues further away from the asymptote would benefit more on reinforced trials and suffer more on non-reinforced trials from the common error term than would cues with less extreme initial associative values. These ideas have been formally expressed and extended by Holmes and colleagues (2019) in a double sigmoid function which maps associative strength non-linearly onto performance. Specifically, the summed error contributes least to the ability to elicit an outcome expectation for cues located around zero and the two horizontal asymptotes (i.e. maximum excitatory and maximum inhibitory strength) and its contribution increases as a cue's associative value moves further away from these parts of the curve. This mapping function has met with considerable success in accounting for the uneven distribution of associative change across members of a compound, but has recently been called into question by a set of results observed in human causal learning (Spicer et al., 2022).

Spicer et al. (2020) articulated the idea that humans exhibit a tendency to preserve existing causal knowledge about relationships between environmental events, and to achieve such a goal, conflicting information encountered in later experience is attributed to the cue that has a less certain causal status. For example, in the case of extinguishing a compound composed of an excitatory cue and a neutral cue, the theory protection principle favors the neutral cue as the one undergoing more associative change than its companion. Since the neutral cue is not paired with any outcome prior to compound training, its causal status remains unknown. By contrast, the causal nature of the excitatory cue can be unambiguously

derived from previous training. The greater uncertainty associated with the neutral cue then makes it more liable to associative decrement than the generative cue present in the same compound. This prediction of the theory protection principle is divergent from those of error-correction algorithms. Individual error term models anticipate less learning about the neutral cue as it is closer to the asymptote of learning supported by the trial outcome compared to the excitor (i.e. the neutral cue generates less prediction error than the generative cause). Common error term models integrating the double sigmoid mapping function envision equal amount of associative change if it is assumed that asymptotic learning has been reached for the excitor prior to compound training. Spicer et al. (2022) observed greater learning about the ambiguous cue than the excitor (as indicated by a difference in the compound test). This observation is counterintuitive because the former conveys a smaller error than the latter, supporting causal ambiguity rather than prediction error as the governing factor for cues involved in compound learning.

The redundancy effect entails the critical comparison between a cue with moderate associative strength but uncertain causal status, the blocked X, and a cue with low associative strength but certain causal status, the uncorrelated Y. This dissociation as demonstrated by Jones and colleagues (2019) forms the important prerequisite for Spicer et al.'s (2020) work. Spicer and colleagues (2020) did not replicate the compound training design with an excitor and a neutral cue or an inhibitor as per Rescorla (2000, 2001), but instead, they took advantage of the dissociative pattern between causal judgment and causal certainty observed for the redundant cues. According to theory protection, if XY+ training follows the standard redundancy effect training, the reluctance to change the strong existing belief that Y is non-causal from previous training should lead to the attribution of the unexpected outcome to the causally ambiguous X, forming the belief that X is causal. As predicted, the authors found that X gained a stronger association with the outcome than did Y during XY compound

learning as indexed by the compound test procedure. Spicer et al.' (2020, 2022) findings suggest that the tendency to update causal knowledge increases in direct proportion to the causal ambiguity of a cue. However, it should be noted that confidence difference in outcome predictions for X and Y was not explicitly tested in these experiments. As the results from Chapter 3 indicate, it cannot be assumed that judgments for Y will be made with greater confidence than judgements for X just because the redundancy effect is present.

Given the cardinal role of causal ambiguity in shaping learning of individual cues in a compound from the perspective of the theory protection principle, it is essential to consider the assumptions that influence the ease with which the causal status of redundant cues can be inferred. For the blocked cue X, magnitude additivity assumptions have been shown to modulate certainty with respect to the causal status of X as well as enhance blocking, presumably because they permit valid deduction to take place (e.g. Beckers et al., 2005; Livesey & Boakes, 2004; Livesey et al., 2019; Lovibond, 2003). Livesey et al. (2019) independently measured confidence and found improved confidence after encouraging additive assumptions compared to when such encouragement was absent. An additive assumption implies that the effects arising from different cues are additive and would sum to a total effect of larger magnitude than that generated by any of the individual cues alone. As X does not increase the magnitude of the effect following AX compound, it would not be regarded as a causal cue under the additive assumption. A non-additive assumption, on the other hand, suggests that the effects caused by individual cues do not sum to a total, and are the same as the effect of multiple cues combined together. The causal status of X under this assumption is ambiguous as either a causal X or a non-causal X would lead to the same outcome on AX trials, making it difficult to discriminate between the two cases.

Although the literature is limited in regards to the effect of prior beliefs on the relative validity effect. Several authors have proposed a similar role for preconceived assumptions on

confidence for the uncorrelated cue Y (Uengoer et al., 2013; Jones et al., 2019). If one holds the general heuristic that valid causes always produce the outcome, Y could easily be ruled out of its potential causal nature on the basis of the intermittent reinforcement schedule it receives, that is, Y leads to outcome occurrence on CY trials but leads to outcome non-occurrence on BY trials. This simple deduction requires an additional assumption that accompanying cues cannot be preventative of an outcome that would otherwise occur. If B is seen as capable of exerting a preventative influence on cooccurring cues, then Y may be considered as a causal cue whose effect on BY trials is prevented by the accompanying B. To summarise, additive assumptions and non-preventative assumptions encourage deductive reasoning for X and Y respectively, and by doing so, enhance learners' confidence that these cues are not valid causes.

Spicer et al. (2020) provides evidence in support of the idea that learners protect existing theory by forming or strengthening the association between the less certain cue and the incompatible outcome, their conclusion, however, is subject to caution for several reasons. Most concerningly, Spicer and colleagues assumed the dissociation between causal judgment and causal certainty for the blocked and uncorrelated cues without measuring learners' confidence. To the best of our knowledge, this dissociative pattern has only been demonstrated in one published study (Jones et al., 2019) and in Chapter 3 of the current thesis where deduction about the uncorrelated cue was strongly encouraged via non-preventative pretraining but not otherwise.

To extend Spicer et al. (2020), the present study introduced a pretraining phase aimed at manipulating certainty with respect to the causal status of redundant cues. Judgments of how likely cues cause the outcome as well as how confident learners are in these judgments were assessed before and after the compound training phase with a ratings test and a forced choice test. Following from our previous work on the role of deductive reasoning in the

redundancy effect in Chapter 3, additive and preventative pretraining and non-additive and non-preventative pretraining were implemented to allow deduction for one of the critical redundant cues while discouraging reasoning for the other. An assumption-neutral pretraining was included to control for the effect of extra training.

Additive and preventative pretraining encourages deductive reasoning about X by highlighting additive rules when integrating effects over multiple causes and discourages deductive reasoning about Y by raising the possibility that the accompanying B may prevent its predicted outcome from occurring. As a consequence, learners should be confident that X cannot be a valid causal cue but be uncertain as to whether Y is causal or non-causal and judge the causal likelihood of Y to be somewhere around the middle. If this obliteration of the redundancy effect occurred then, according to the theory protection principle, it should have consequences for the individual learning about X and Y when both cues are subsequently presented in compound. The presence of the outcome on Stage 2 XY trials should now be inconsistent with the more certain knowledge that X is non-causal. Instead, the outcome should be attributed to Y, which now has an ambiguous causal status, in an attempt to preserve the causal knowledge acquired about X in Stage 1.

On the other hand, non-additive and non-preventative pretraining makes deduction invalid for X but fulfills conditions necessary to draw deductive inferences for Y. This has the consequence of reducing confidence for X which can either be causal or non-causal under a non-additive rule, but elevating confidence for Y as its outcome cannot be prevented by the co-present B on outcome absent trials if preventative relationships are considered plausible. The redundancy effect should therefore be enhanced following non-additive and non-preventative pretraining. If learners engaged in theory protection of the form proposed by Spicer and colleagues (2020), then there should be stronger updating of causal knowledge about Y than about X when they are subsequently reinforced together. This is because the

presence of the outcome on XY trials is in conflict with the prediction of its absence elicited by Y but is less so with the prediction elicited by X. Learners therefore establish a stronger association between X and the outcome than between Y and the outcome as a means to minimise violation of existing knowledge that Y is non-causal.

Experiment 4.1

The theory protection principle suggests that, when faced with incompatible new information, more learning will occur about less certain cues. Experiment 4.1 sought to examine this account by directly manipulating the level of certainty with regard to judgments about the blocked and uncorrelated cues before introducing the compound test procedure in accordance with Spicer et al. (2020). This was done by instilling participants with either additive and preventative assumptions or non-additive and non-preventative assumptions in a pretraining phase. The former would encourage the deduction that the blocked cue cannot be causal with high confidence but make the causal status of the uncorrelated cue ambiguous, while the latter would maintain uncertainty to the blocked cue but enhance the simple inference that the uncorrelated cue must not be causal. In both cases, certainty that learners possess in making likelihood judgments should dissociate from the strength of the causal judgements themselves. The theory protection principle assumes that people will attempt to retain the causal knowledge about cues which they are most confident. In the additive and preventative group, this should be knowledge that X is not causal, meaning more should be learned about Y. In the non-additive and non-preventative group, this should be knowledge that Y is not causal, meaning more will be learned about X. Empirical confirmation of this result is critical for distinguishing theory protection from alternative error-correction accounts as none of the associative accounts speaks to changes in causal certainty as a result of prior experience and its influence on subsequent compound learning of redundant cues.

Method

Participants

One hundred and twenty-three first-year psychology students from the University of Sydney participated in this experiment in return for course credits. The sample size was chosen based on Spicer et al. (2020; Experiment 1) where 36 participants were recruited for a single group experiment. Participants were randomly allocated to one of three groups: the additive and preventative group, the non-additive and non-preventative group, or the neutral group. In keeping with the conventional performance criterion in Chapter 3 and other causal learning research (e.g. Don & Livesey, 2017; Livesey et al., 2019), participants who failed to pass the learning threshold of 60% in any training Stages were removed from further analyses. Average accuracy in the last half of pretraining and main training was compared against the learning criterion, whereas overall accuracy was used for compound training because of its brevity. This criterion resulted in the exclusion of 4 participants, leaving 39 (28 females, mean age=19.36, *SD*=1.55) in the additive and preventative group, 40 (34 females, mean age=19.40, *SD*=3.00) in the non-additive and non-preventative group, and 40 (28 female, mean age=19.08, *SD*=1.05) in the control group. All three groups have over 86% power to detect any difference in learning between the blocked and uncorrelated cues during compound training of a similar magnitude reported in Spicer et al. (2020, Experiment 1; *dz* = .50).

Table 4.1

The design of Experiment 4.1

Condition	Pretraining	Stage 1 Training	Stage 1 Test	Stage 2 Training	Stage 2 Test	Cue choice test
neutral	M+ N+ O– P–	Blocking: A+ AX+		XY+	XZ, WY	X vs. Y
additive & preventative	M+ N+ O– P– MN++ OP– MO– NP–	D+ DW+ Relative Validity: BY– CY+ EZ– FZ+	X, W, Y, Z, A, D, B, E, C, F		X, W, Y, Z A, D, B, E, C, F	X vs. Z W vs. Y W vs. Z X vs. W Y vs. Z
non-additive & non-preventative	M+ N+ O– P– MN+ OP– MO+ NP+					

Note. Letters represent cues, randomly assigned to different medicine names. "+" represents hormone increase and "–" represents no hormone change.

Design

The design of Experiment 4.1 is illustrated in Table 4.1. There were three training phases comprising 4 blocks of pretraining, 8 blocks of Stage 1 training and 3 blocks of Stage 2 training. Within each training block, each trial type was presented twice with counterbalanced spatial location and in randomised trial order. A test phase was conducted before and after the interim compound training, assessing both perceived likelihood that the outcome would follow and the level of confidence in these likelihood judgments. The key test compounds XZ and WY each consisted of one blocked cue and one uncorrelated cue from Stage 1 training. The only aspect that they differed was that the former had the blocked cue X while the latter had the uncorrelated cue Y undergo compound training. Any difference in associative update for X and Y during compounded learning should thus be captured by the judgment difference between XZ and WY. Each test phase involved two presentations of each test trial to counterbalance cue positions. Responses in both tests were measured in the form of ratings and were additionally recorded as forced choices from cue pairs on the final test. The use of an interim test was to address an obvious limitation in Spicer et al. (2020)

where causal judgment of the critical cues was not assessed until the end of compound training. For the pair that underwent additional training, it would be difficult for associations acquired in Stage 1 to survive the interference from compound training. For the pair that did not, there could still be possible disruptions due to subsequent learning even though they had no direct involvement. The hormone change task developed by Zaksaite and Jones (2020) was adopted as preventative relationships between medicines and hormone change are more natural and malleable in the real world than between foods and ailment in the traditional food allergist task.

Procedure

The apparatus and stimuli were the same as per Chapter 3. The experimental cues were fictitious medicine names each beginning with a unique letter from A to R. These included *Aspetur, Broncin, Chrurin, Dioxnyl, Ephemerol, Felicium, Gambutrol, Hyronalin, Impbatine, Jamitol, Krayoxx, Lithorol, Metazine, Nozambutol, Ontapelium, Plycidox, Quelinum, and Rodvuccial*. The first experiment was based off Experiment 1 of Spicer et al. (2020) with an added pretraining phase. The additive and preventative pretraining aimed to enhance confidence for X but reduce confidence for Y, while the non-additive and non-preventative pretraining aimed to lower confidence for X but elevate confidence for Y. The additive and preventative pretraining group and the non-additive and non-preventative pretraining group also received detailed instructions in written form regarding the targeted assumptions.

All groups received pretraining on four cues (i.e. M, N, O, P) that were not used in the later training stages. During pretraining, the additive and preventative participants also made two important observations. First, when two medicines that both led to an increase in hormone levels were taken simultaneously, the resultant hormone increase was larger than the increase caused by the individual medicines on their own. Second, when one medicine paired with hormone increase and another medicine with no impact on hormone levels were

taken together, the effect on hormone levels of the generative cause was prevented by the other medicine consumed at the same time. The non-additive and non-preventative group learned two opposite rules. First, when two effective medicines were administered at the same time, the combined effect was the same as that of a single effective medicine. Second, when an effective medicine and an ineffective medicine were taken simultaneously, the hormone increase caused by the effective medicine remained unchanged.

Without further instructions or break, participants proceeded to Stage 1 training which involved a set of blocking trials and a set of relative validity trials. An interim test was conducted after Stage 1 training, requiring participants to rate the likelihood that individual or compounded cues were causes of the outcome and then rate the degree of confidence they possessed in their ability to make these judgments. Participants then proceeded to Stage 2 compound training where they were told that they would observe more medicines that Patient X consumed and be asked to predict their effects. During Stage 2 training, the blocked cue and the uncorrelated cue were simultaneously trained in the presence of the outcome. Participants then completed a second ratings test that was the same as the first. After this, participants completed a forced choice test where participants chose the cue that was perceived to be the more likely cause and then chose the cue that they felt more confident judging regardless of whether the cue was believed to be a cause or a non-cause from each cue pair. The choice test was kept at the end for consistency with previous chapters but since it did not test the critical compounds and was not used for any of the main conclusions, the results will be reported in supplementary materials only.

Results

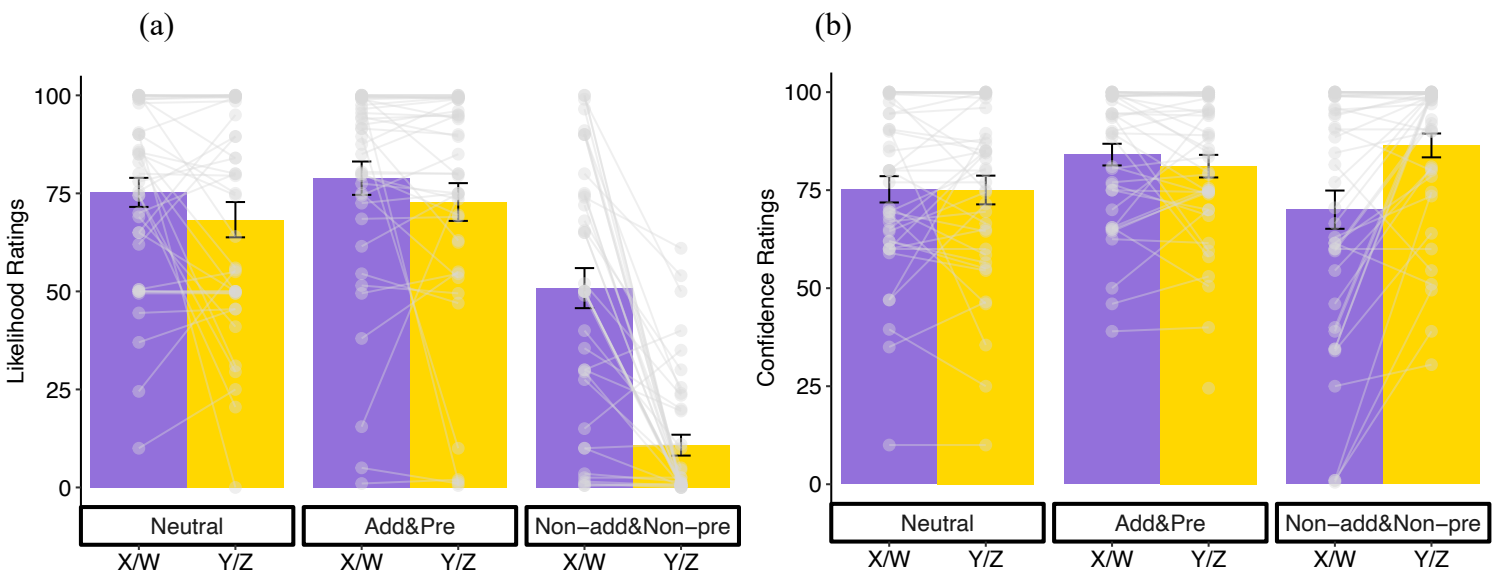
Training

Cue-outcome contingencies were learned rapidly in all three groups. The additive and preventative group reached an average accuracy of 98.40% in the last half of pretraining,

98.32% in the last half of main training, and 93.39% in compound training; the non-additive and non-preventative group reached a similar accuracy of 97.73% in the last half of pretraining, 97.11% in the last half of main training, and 90.42% in compound training; the neutral group achieved a slightly lower accuracy of 95.16% in the last of pretraining, 96.13% in the last half of main training, and 97.08% in compound training. Accuracy was assessed with a mixed-model ANOVA with block as the within-subjects factor, pretraining group as the between-subjects factor, and average correct predictions in each block as the dependent variable. In comparison to the neutral group, the quadratic trend for block was stronger for the additive and preventative group in pretraining, $t(78)=2.03$, $p=.044$, $d=.228$, and main training, $t(78)=3.12$, $p=.002$, $d=.351$, and stronger for the non-additive and non-preventative group in pretraining, $t(79)=2.31$, $p=.023$, $d=.258$, and main training, $t(79)=4.00$, $p<.001$, $d=.447$. These results suggest that participants in the meaningful pretraining conditions have learned the cue-outcome contingencies faster and reached ceiling earlier during pretraining and main training than those in the control. See supplementary material for further details on training performance.

Figure 4.1

(a) Mean likelihood ratings and (b) Mean confidence ratings on the interim ratings test for the neutral pretraining group, the additive and preventative group, and the non-additive and non-preventative group in Experiment 4.1. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.



Interim Test

Likelihood. Likelihood and confidence ratings given to individual cues and cue compounds at the interim test are illustrated in Figure 4.1. Please see supplementary materials for complete figures including all cues for Experiments 4.1-4.4. The effect of most interest was the redundancy effect as reflected in the comparison between the average of the two blocked cues X and W and the average of the two uncorrelated cues Y and Z. To examine the effect of prior assumptions on the redundancy effect, a (2)x3 mixed-measures ANOVA with cue type (blocked vs. uncorrelated) as the within-subjects factor, pretraining condition (neutral vs. preventative and additive vs. non-preventative and non-additive) as the between-subjects factor, and likelihood ratings as the dependent variable was run. Results showed that the blocked cues (i.e. X & W) were given higher likelihood ratings than the uncorrelated cues (i.e. Y & Z) averaged across pretraining group, $F(1,116)=47.40$, $p<.001$, $\eta_p^2=.290$, indicating the presence of an overall redundancy effect. The average ratings to cues were found to differ substantially across groups, $F(2,116)=48.04$, $p<.001$, $\eta_p^2=.453$. Importantly, the magnitude of the blocked-uncorrelated difference changed significantly depending on the pretraining received, $F(2,116)=19.60$, $p<.001$, $\eta_p^2=.253$, suggesting a significant effect of pretraining condition on the redundancy effect.

Pairwise comparisons between the two meaningful pretraining groups and the neutral group revealed that the higher average likelihood rating to the blocked cues than to the uncorrelated cues was more marked in the non-additive and non-preventative group compared to the neutral group, $t(78)=4.72$, $p=.014$, $d=1.056$, but was similar between the additive and preventative group and the neutral group, $t(77)=.87$, $p=.387$, $d=.196$. Subsequent simple contrasts showed that the redundancy effect was present in the neutral group, $t(39)=2.22$, $p=.032$, $d=.352$, and the non-additive and non-preventative group, $t(39)=7.53$, $p<.001$, $d=1.191$, but was no longer significant in the additive and preventative group,

$t(38)=1.04$, $p=.305$, $d=.166$. Additional Bayesian binomial tests revealed anecdotal evidence for the redundancy effect for the neutral group, $BF_{10}=1.536$, and evidence in favour of the null for the additive and preventative group, $BF_{10}=.285$.

To evaluate the effect of target assumptions on respective redundant cues, ratings to the blocked and uncorrelated cues were compared between the two meaningful pretraining conditions. While additive and preventative pretraining substantially elevated likelihood ratings to the uncorrelated cues relative to non-additive and non-preventative pretraining, $t(77)=11.95$, $p<.001$, $d=2.689$, the blocked cues were rated as more likely causes by additive and preventative participants compared to non-additive and non-preventative participants, $t(77)=4.21$, $p<.001$, $d=.947$.

Confidence. Interim confidence ratings were analysed with the same mixed-measures ANOVA but with confidence as the dependent variable. Results showed that confidence was significantly higher for the uncorrelated cues than for the blocked cues averaged over group, $F(1,116)=7.10$, $p=.009$, $\eta_p^2=.058$, and this confidence difference between critical cues varied across pretraining condition, $F(2,116)=12.17$, $p<.001$, $\eta_p^2=.173$. The main effect of group was however non-significant, $F(2,116)=1.88$, $p=.158$, $\eta_p^2=.031$. Between-group comparisons revealed that non-additive and non-preventative pretraining enhanced confidence for the uncorrelated cues relative to the blocked cues in comparison to neutral pretraining, $t(78)=3.76$, $p<.001$, $d=.840$. Although inspection of Figure 4.1b indicates that additive and preventative pretraining numerically improved confidence for the blocked cues relative to the uncorrelated cues compared to neutral pretraining, this difference was not statistically significant, $t(77)=.113$, $p=.910$, $d=.025$.

Further simple contrasts confirmed that the participants only felt more confident making likelihood judgments about the uncorrelated cues than about the blocked cues following non-additive and non-preventative pretraining, $t(39)=4.26$, $p<.001$, $d=.674$, but

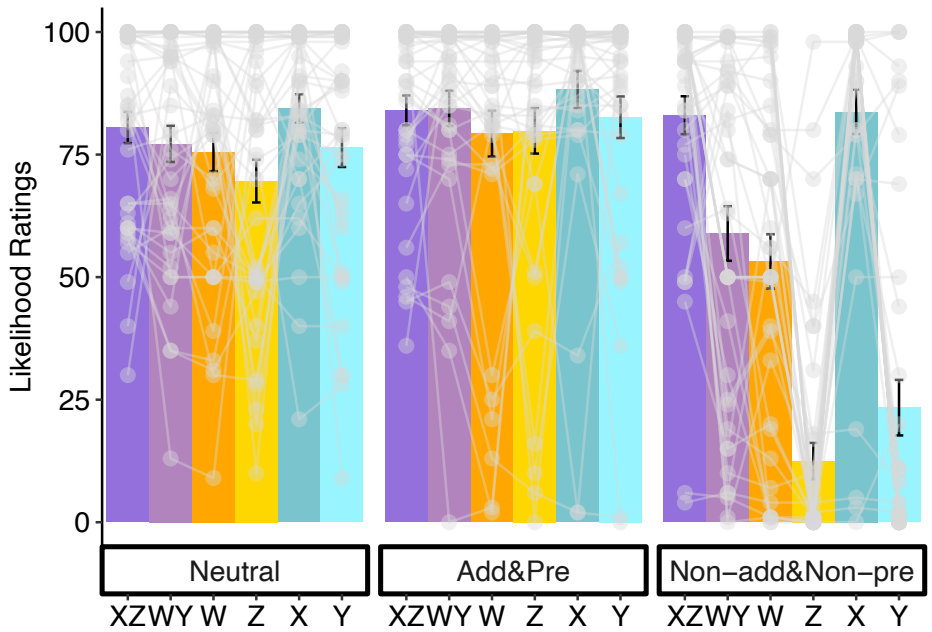
there was no confidence difference after additive and preventative pretraining $t(38)=.80$, $p=.430$, $d=.128$, or neutral pretraining, $t(39)=.45$, $p=.656$, $d=.071$. Additional Bayesian binomial test revealed evidence in favour of the null for the additive and preventative group, $BF_{10}=.232$, and the neutral group, $BF_{10}=.188$. The effect of target assumptions on confidence was assessed for each redundant cue between the two meaningful pretraining groups through two pairwise comparisons. Results revealed higher confidence for the blocked cues following additive and preventative pretraining than following non-additive and non-preventative pretraining, $t(77)=2.49$, $p=.015$, $d=.947$, but no confidence difference was detected for the uncorrelated cues between the two assumption-biased pretraining conditions $t(77)=.97$, $p=.334$, $d=.226$.

Final Test

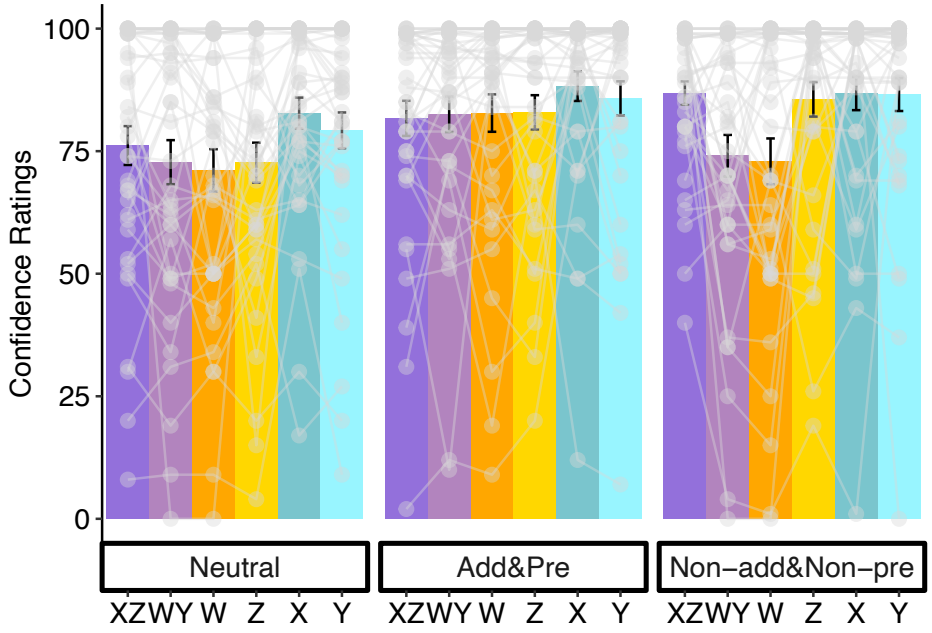
Figure 4.2

(a) Mean likelihood ratings and (b) Mean confidence ratings on the final ratings test for the neutral pretraining group, the additive and preventative group, and the non-additive and non-preventative group in Experiment 4.1. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

(a)



(b)



Ratings Test. Likelihood and confidence ratings given on the final test are illustrated in Figure 4.2. Two main comparisons were of interest. The first was between compounds XZ and WY to examine any difference in Stage 2 compound training about the blocked cue X and the uncorrelated cue Y. A (2)x3 mixed model ANOVA with compound type (XZ vs. WY) as the within-subjects factor and pretraining group as the between-subjects factor was run on likelihood ratings and then on confidence ratings to achieve this purpose. The second was between the critical redundant cue pairs, which would reflect an influence of compound training on the redundancy effect. That is, the blocked vs. uncorrelated cue difference was compared between the pair of cues X and Y that underwent compound training and the pair of cues W and Z that did not. A (2)x(2)x3 mixed model ANOVA with cue type (blocked vs. uncorrelated) and compound training (additionally trained vs. not additionally trained) as within-subjects factors and pretraining group as the between-subjects factor was run on likelihood ratings and then on confidence ratings to interrogate this question.

Critical Compounds Likelihood. For the first comparison, the ANOVA test with likelihood ratings as the dependent variable revealed that XZ was rated as a more likely cause

of the outcome than WY averaged over group, $F(1,116)=12.16$, $p<.001$, $\eta_p^2=.095$. Ratings to these compounds on average varied across groups, $F(2,116)=4.28$, $p=.016$, $\eta_p^2=.069$. As expected, the higher likelihood ratings to the compound containing X over the compound containing Y depended on the pretraining received, $F(2,116)=8.68$, $p<.001$, $\eta_p^2=.130$. This interaction was subsequently examined in pairwise comparisons between groups. Results revealed that the greater learning about X than about Y was more marked following non-additive and non-preventative pretraining than following neutral pretraining, $t(77)=3.55$, $p<.001$, $d=.799$. This learning difference for X and Y differed significantly between the additive and preventative group and the non-additive and non-preventative group, $t(78)=2.97$, $p=.004$, $d=.665$, but did not vary between the additive and preventative group and the neutral group, $t(77)=.76$, $p=.448$, $d=.171$. Further simple contrasts revealed significantly higher ratings to XZ than to WY among non-additive and non-preventative participants, $t(39)=4.02$, $p<.001$, $d=.636$, but no such difference was revealed among additive and preventative participants, $t(38)=.12$, $p=.909$, $d=.019$. Although XZ was given slightly higher ratings than WY in the neutral group, this difference failed to reach statistical significance, $t(39)=.93$, $p=.357$, $d=.147$.

Critical Compounds Confidence. Confidence for XZ and WY were analysed through the same ANOVA but with confidence ratings as the dependent variable. There was a main effect of compound, $F(1,116)=7.04$, $p=.009$, $\eta_p^2=.057$, indicating that participants were more certain about the causal status of XZ than about WY averaged over groups. However, this greater degree of certainty associated with likelihood judgments for XZ than for WY differed among the groups, $F(2,116)=4.34$, $p=.015$, $\eta_p^2=.070$. The group main effect was not significant, $F(2,116)=1.41$, $p=.248$, $\eta_p^2=.024$. The interaction between compound and group was investigated further in simple contrasts comparing the size of this confidence difference across groups. Results showed that the confidence difference for XZ and WY differed

significantly between the additive and preventative group and the non-additive and non-preventative group, $t(77)=2.84$, $p=.006$, $d=.639$, but did not between either the additive and preventative group and the neutral group, , $t(77)=1.05$, $p=.295$, $d=.237$, or the non-additive and non-preventative group and the neutral group, $t(78)=1.78$, $p=.079$, $d=.398$. Simple t-tests confirmed that participants were only more confident about their likelihood judgments for XZ over YZ after non-additive and non-preventative pretraining, $t(39)=3.07$, $p=.004$, $d=.486$, but were not after additive and preventative pretraining, $t(38)=.36$, $p=.724$, $d=.057$, or neutral pretraining, $t(39)=1.05$, $p=.302$, $d=.165$.

Critical Cues Likelihood. The ANOVA for the second likelihood ratings comparison revealed a significant interaction between cue pair and compound training, indicating larger redundancy effect between X and Y than between W and Z averaged over group, $F(1,116)=5.28$, $p=.023$, $\eta_p^2=.044$. The magnitude of the redundancy effect on average differed significantly across groups, $F(2,116)=40.86$, $p<.001$, $\eta_p^2=.413$. The three-way interaction between cue pair, compound training, and group was however not significant, $F(2,116)=1.72$, $p=.184$, $\eta_p^2=.029$. Inspection of Figure 4.2a indicates that ratings were higher for X than for Y and higher for W than for Z in the non-additive and non-preventative group compared to the other groups. Independent samples t-tests confirmed that the higher likelihood ratings to X than to Y was more marked following non-additive and non-preventative pretraining than following additive and preventative pretraining $t(77)=6.38$, $p<.001$, $d=1.436$, or neutral pretraining, $t(78)=6.57$, $p<.001$, $d=1.468$. Paired samples t-tests revealed that the redundancy effect was significant between X and Y in the non-additive and non-preventative group, $t(39)=8.31$, $p<.001$, $d=1.313$, and the neutral group, $t(39)=2.42$, $p=.020$, $d=.383$, but was eliminated in the additive and preventative group, $t(38)=1.27$, $p=.211$, $d=.204$, $BF_{10}=.365$. Likewise, the higher ratings to W over Z were significantly more pronounced among non-additive and non-preventative participants compared to neutral

pretraining participants, $t(78)=4.41$, $p<.001$, $d=.986$. The ratings difference between W and Z also varied significantly between the non-additive and non-preventative participants and the additive and preventative participants, $t(77)=5.04$, $p<.001$, $d=1.135$. Additive and preventative pretraining appears to have slightly reduced the ratings difference between W and Z compared to the neutral pretraining group, this attenuation however did not reach significance, $t(77)=.860$, $p=.392$, $d=.194$. Additional simple t-tests showed that the redundancy effect in the form of higher likelihood ratings to W over Z was only present in the non-additive and non-preventative group, $t(39)=6.79$, $p<.001$, $d=1.073$, but was not present in either the additive and preventative group, $t(38)=.11$, $p=.916$, $d=.017$, $BF_{10}=.174$, or the neutral pretraining group, $t(39)=1.154$, $p=.256$, $d=.182$, $BF_{10}=.316$.

Critical Cues Confidence. The ANOVA for the second confidence ratings comparison revealed a significant interaction between cue type and compound training, suggesting that whether confidence ratings were higher for the blocked cue over the uncorrelated cue averaged over group depended on whether the cue pair underwent compound training, $F(1,116)=6.50$, $p=.012$, $\eta_p^2=.053$. There was however no significant interaction between cue type and group, $F(2,116)=2.97$, $p=.055$, $\eta_p^2=.049$, or among cue type, compound training, and group, $F(2,116)=1.30$, $p=.277$, $\eta_p^2=.022$. Further independent samples t-tests indicated that the confidence difference between Z and W did not differ between the additive and preventative group and the neutral group, $t(77)=.29$, $p=.773$, $d=.065$ or between the non-additive and non-preventative group and the neutral group, $t(78)=1.78$, $p=.080$, $d=.397$. Simple t-tests revealed that participants only felt more confident about Z than W in the non-additive and non-preventative condition, $t(39)=2.81$, $p=.008$, $d=.444$, but not in either the additive and preventative condition, $t(38)=.07$, $p=.946$, $d=.011$, or the neutral pretraining condition, $t(39)=.37$, $p=.717$, $d=.058$.

Discussion

Experiment 4.1 found evidence of theory protection in compound training when pretrained assumptions facilitated a clear dissociative pattern between likelihood judgment and causal certainty. This pattern was particularly evident in the non-additive and non-preventative pretraining condition, where learners were confident that the uncorrelated cue is non-causal but uncertain about the causal role of the blocked cue. In this group, there was strong evidence that the blocked cue X received greater updating than the uncorrelated cue Y, during XY compound training. In line with the proposal of Spicer et al. (2020), this result suggests that in order to maintain the existing knowledge that Y is not a valid cause, participants attributed the outcome on compound XY trials more to the causally ambiguous X than they did to the certainly non-causal Y. Participants given assumption-neutral prior training regarded the blocked cues as more likely causes than the uncorrelated cues but did not display a confidence pattern that opposed likelihood judgment. The neutral group subsequently revealed no evidence of any learning difference in Stage 2 compound training. This failure to find theory protection implies a critical role for differential initial confidence in promoting unequal learning about concurrently trained redundant cues, while the presence of the judgment difference (i.e. the redundancy effect) alone is insufficient. Results from the additive and preventative group also demonstrated *some* consistency with theory protection. Encouraging deductive reasoning for the blocked cues while discouraging its use for the uncorrelated cues did not fully reverse the redundancy effect and the associated confidence pattern, but reduced both judgment differences to non-significance. The lack of asymmetry in subsequent updating during Stage 2 training is at least consistent with the prediction of theory protection in that if one possessed equal confidence in their judgments for X and Y, both cues should be updated to the same extent during later training as a compound. It should be noted that the ratings for the blocked cues were much higher in this than we expected given the

additive assumptions that we tried to implement in pretraining. It is possible that concurrent pretraining of preventative assumptions was unintentionally responsible for this, leading to no change in the ambiguous causal status of the blocked cues but an increase in uncertainty around the uncorrelated cues. On the whole, these findings suggest that learning about constituent cues of a compound is not merely determined by the starting associative strengths of component cues, but is also influenced by the initial certainty with respect to the causal status of these cues.

There are two aspects of the Experiment 4.1 design that might have inadvertently led to persistently high Stage 1 test ratings for the blocked cues, thus reducing the efficacy of the additivity pretraining. First, in combining additivity and preventative pretraining, Experiment 4.1 presented two examples of compounds that combined a causal cue and a non-causal cue, and led to no outcome occurring (i.e. M+ / O– / MO– and N+ / P– / NP–). This may have led participants in this group to expect non-causal cues to usually be preventative, rather than merely encouraging them to consider prevention plausible. In doing so, we may have inflated the judgment that X was causal because it did not prevent the outcome on AX+ trials. Second, since we did not include filler trials in Stage 1 or Stage 2 learning, there was a high probability of the outcome occurring throughout most of the task. This could potentially reduce blocking and inflate causal ratings for the blocked cues (Jones et al., 2019).

Using a design that addressed these potentially issues, Experiments 4.2–4.4 sought to replicate the differences between additive and preventative vs. non-additive and non-preventative pretraining (Experiment 4.2), then examine effects of additivity pretraining (Experiment 4.3) and preventative pretraining (Experiment 4.4) independently. Since these experiments were identical in all respects except the pretraining, they are presented together below.

Experiments 4.2 to 4.4

The design of the next three experiments differed from Experiment 4.1 in terms of the pretraining, as well as the addition of filler cues to Stage 1 and Stage 2 training in order to maintain a more even split of trials that resulted in hormone change and trials that resulted in no change. As can be seen in Table 4.2, the three experiments used identical designs for Stage 1 and Stage 2 training, and all test phases were also the same. The three experiments thus only differed in terms of the type of pretraining delivered, and the causal instructions that accompanied that pretraining.

Partially replicating Experiment 4.1 (though with the omission of the neutral pretraining group), Experiment 4.2 compared a non-additive and non-preventative pretraining condition to an additive and preventative condition. This time, the preventative pretraining component demonstrated that prevention by non-causal cues only occurred *sometimes* by showing a compound in which the outcome was prevented (M+ / O– / MO–) and another where it was not prevented (N+ / P– / NP+). We assumed that if Stage 1 ratings for the blocked cues W and X were still unexpectedly high after these modifications then it suggests additivity pretraining was ineffective for other reasons, perhaps to do with the combination of additivity and preventative assumptions. But regardless, this provided an attempt to replicate the key result in Experiment 4.1 with an improved design. Experiments 4.3 and 4.4 then manipulated the additive/non-additive and preventative/non-preventative pretraining components independently to assess their individual contributions.

Table 4.2

Design of Experiments 4.2-4.4

Experiment	Condition	Pretraining	Stage 1 Training	Stage 1 Test	Stage 2 Training	Stage 2 Test	Cue choice test
4.2	additive & preventative	M+ N+ O– P– MN++ OP– MO– NP+	Blocking: A+ AX+ D+ DW+	X, W, Y, Z	XY+	XZ, WY	X vs. Y
	non-additive & non-preventative	M+ N+ O– P– MN+ OP– MO+ NP+	Relative Validity: BY– CY+ EZ– FZ+	A, D, B, E, C, F	QR–	X, W, Y, Z A, D, B, E, C, F	X vs. Z W vs. Y W vs. Z X vs. W Y vs. Z
4.3	additive	M+ N+ O– P– MN++ OP–	Fillers: G– H– IJ– KL–				
	non-additive	M+ N+ O– P– MN+ OP–					
4.4	preventative	M+ N+ O– P– MO– NP+					
	non-preventative	M+ N+ O– P– MO+ NP+					

Note. Letters represent cues, randomly assigned to different medicine names. "+" represents hormone increase and "–" represents no hormone change.

Method

Participants and Apparatus

For all three experiments, undergraduate psychology students from the University of Sydney were recruited (N=111 for Experiment 4.2, N=103 for Experiment 4.3, and N = 117 for Experiment 4.4). They participated in person in return for course credit, under the same conditions and using the same apparatus as Experiment 4.1, allocated to condition within each experiment according to time of arrival.

In Experiment 4.2, eight participants performed below the 60% criteria in one or more of the training Stages, and were removed from further analyses. The remaining sample comprised 52 (40 females, mean age =19.87, *SD*=4.84) in the additive and preventative group, and 51 (37 females, mean age =19.67, *SD*=3.05) in the non-additive and non-preventative group.

In Experiment 4.3, six participants were excluded for failing to pass the learning criteria. This resulted in 48 (36 females, mean age =19.92, *SD*=2.58) in the additive group and 49 (28 females, mean age =19.67, *SD*=1.53) in the non-additive group.

In Experiment 4.4, 10 participants were excluded for failing the learning criteria, leaving 50 (37 females, mean age =19.16, *SD*=1.50) in the preventative group, and 57 in the non-preventative group (45 females, mean age =19.23, *SD*=2.31).

Design and Procedure

Experiment 4.2 focused on the comparison between additive and preventative pretraining versus non-additive and non-preventative pretraining with an improved design as shown in Table 4.2. With the aim of alleviating the unintended emphasis on preventative assumptions, the generative effect of N which was prevented by the accompanying non-causal P in the Experiment 4.1 remained unchanged in the current experiment. Additionally, outcome-absent filler trials, including an outcome-absent QR trial in Stage 2, were added

balance out the number of trials experienced as reinforced and non-reinforced. The rest of the design was the same as per Experiment 4.1.

In Experiment 4.3, outcome additivity assumptions, necessary to deduce the non-causality of the blocked cues, were manipulated in pretraining. Participants were randomly assigned to either the additive group where instructions and observed causal relationships encouraged additive assumptions, or the non-additive group where magnitude non-additivity was emphasised in instructions and pretraining. Prior knowledge about possible outcome prevention, necessary for deduction about the uncorrelated cues, was not manipulated. All participants received standard instructions that mention the possibility of cues preventing hormone increase, but these instructions were not given additional emphasis and were not illustrated directly with compound cues during pretraining.

In Experiment 4.4 pretraining manipulated participants' assumptions regarding preventative cue-outcome relationships which were shown to be plausible in the preventative group and implausible in the non-preventative group, using the same instructions and compounds as Experiment 4.2. Additivity assumptions were not manipulated; no reference to additivity or non-additivity was made in instructions and the pretraining compounds did not direct illustrate additivity or non-additivity.

Results

Training

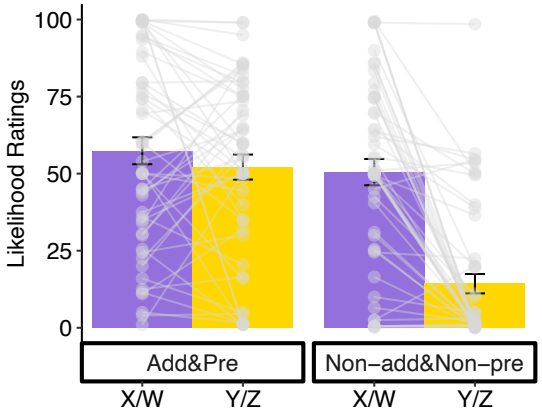
Details about training accuracy for Experiments 4.2–4.4 are reported in supplementary materials. Across all three experiments, cue-outcome relationships were generally learned rapidly across blocks in all three training phases. Performance in the second half of pretraining was generally very accurate (Experiment 4.2, additive and preventative: 93.21%; Experiment 4.2, non-additive and non-preventative: 96.26%; Experiment 4.3, additive: 98.09%; Experiment 4.3, non-additive: 95.66%; Experiment 4.4, preventative: 80.79%;

Experiment 4.4 non-preventative: 89.63%). Likewise, across the second half of Stage 1 training, accuracy was generally high (Experiment 4.2, additive and preventative: 94.53%; Experiment 4.2, non-additive and non-preventative: 95.96%; Experiment 4.3, additive: 94.29%; Experiment 4.3, non-additive: 92.81%; Experiment 4.4, preventative: 86.45%; Experiment 4.4 non-preventative: 86.84%), as was accuracy averaged across all of the Stage 2 training (Experiment 4.2, additive and preventative: 91.03%; Experiment 4.2, non-additive and non-preventative: 90.20%; Experiment 4.3, additive: 90.10%; Experiment 4.3, non-additive: 89.97%; Experiment 4.4, preventative: 93.00 %; Experiment 4.4 non-preventative: 90.64%).

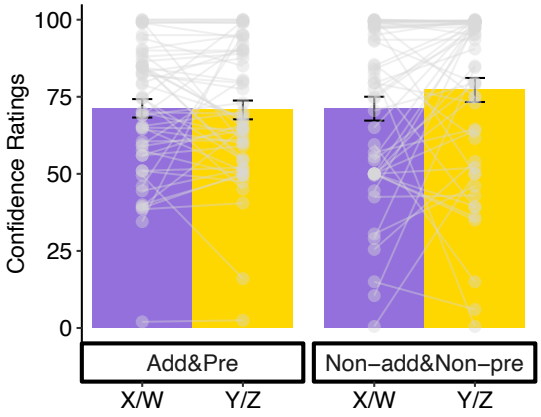
Figure 4.3

Mean likelihood ratings and mean confidence ratings for the critical cues on the interim ratings test for Experiments 4.2-4.4. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

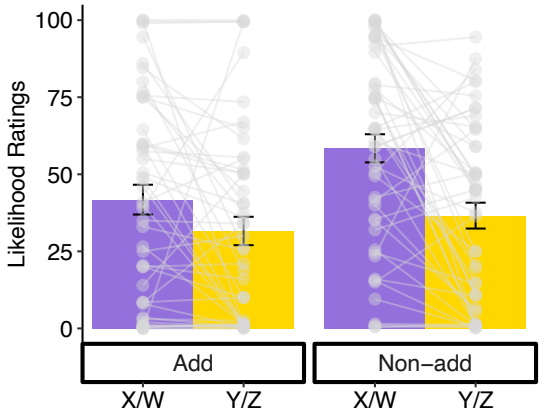
(a) Experiment 4.2 Likelihood Rating



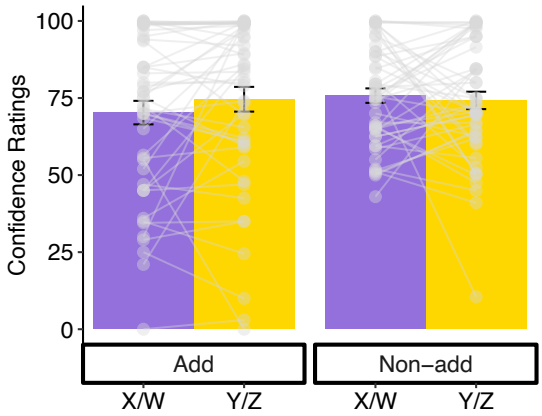
(b) Experiment 4.2 Confidence Rating



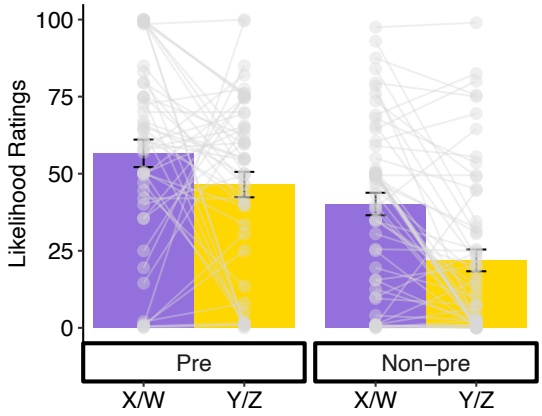
(c) Experiment 4.3 Likelihood Rating



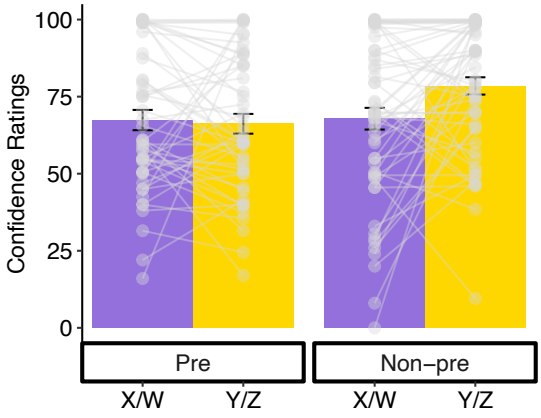
(d) Experiment 4.3 Confidence Rating



(e) Experiment 4.4 Likelihood Rating



(f) Experiment 4.4 Confidence Rating



Stage 1 Interim Test

Figure 4.3 illustrates the likelihood and confidence ratings given to individual blocked cues (X / W) and uncorrelated cues (Y / Z) in each of Experiments 4.2-4.4. Here, statistical results from each rating variable in experiment were analysed using (2)x2 mixed design ANOVA with cue type as the within-subjects factor and pretraining condition as the between-subjects factor. Results for each experiment are reported in sequence.

Experiment 4.2 Likelihood. On average, participants gave higher likelihood ratings to the blocked cues (i.e. X & W) than to the uncorrelated cues (i.e. Y & Z), $F(1,101)=43.19$, $p<.001$, $\eta_p^2=.300$, indicating an overall redundancy effect. On average, additive and preventative participants rated the critical cues as being more likely causes than non-additive and non-preventative participants, $F(1,101)=22.85$, $p<.001$, $\eta_p^2=.185$. Of most interest, the tendency to rate the blocked cues as being more likely causes than the uncorrelated cues was greater in the group that received non-additive and non-preventative pretraining compared to the group that received additive and preventative pretraining, $F(1,101)=23.99$, $p<.001$, $\eta_p^2=.192$. Further exploration of the cue type x group interaction with paired samples t-tests showed that the redundancy effect was significant in the non-additive and non-preventative group, $t(50)=7.95$, $p<.001$, $d=1.113$, but was not present in the additive and preventative group, $t(51)=1.21$, $p=.233$, $d=.167$, $BF_{10}=.300$.

Experiment 4.2 Confidence. The same analysis run on confidence ratings revealed non-significant main effects of cue type, $F(1,101)=1.39$, $p=.242$, $\eta_p^2=.014$, group, $F(1,101)=.54$, $p=.462$, $\eta_p^2=.005$, and a non-significant interaction, $F(1,101)=1.97$, $p=.164$, $\eta_p^2=.019$. Inspection of Figure 4.3b indicates that non-additive and non-preventative participants were numerically more certain about the causal status of the uncorrelated cues than for the blocked cues, but additional t-test showed that this higher confidence was not statistically significant, $t(50)=1.53$, $p=.133$, $d=.214$.

Experiment 4.3 Likelihood. Average likelihood ratings were significantly higher to the blocked cues (i.e. X & W) than the uncorrelated cues (i.e. Y & Z), $F(1,95)=23.38$, $p<.001$, $\eta_p^2=.198$. Although the strength of the redundancy effect did not differ significantly between groups, $F(1,95)=3.11$, $p=.081$, $\eta_p^2=.032$, inspection of Figure 4.3c suggests that the learning bias toward the blocked cues was greater following non-additive pretraining than following additive pretraining. Simple t-tests indicated a significant redundancy effect in both the additive group, $t(47)=2.24$, $p=.030$, $d=.323$, and the non-additive group, $t(48)=4.55$, $p<.001$, $d=.650$. Importantly, however, likelihood ratings for the blocked cues were lower following additive than non-additive pretraining, $t(95)=2.51$, $p=.014$, $d=.509$, and were similar for the uncorrelated cues between groups, $t(95)=.80$, $p=.425$, $d=.204$.

Experiment 4.3 Confidence. The same ANOVA with confidence ratings as the dependent variable revealed no significant difference between the blocked and uncorrelated cues, $F(1,95)=.41$, $p=.525$, $\eta_p^2=.004$, despite the presence of a significant redundancy effect in both groups. Neither the main effect of group, $F(1,95)=.38$, $p=.537$, $\eta_p^2=.004$, nor the cue type x group interaction, $F(1,95)=1.82$, $p=.180$, $\eta_p^2=.019$, were significant.

Experiment 4.4 Likelihood. On average, ratings to the blocked cues (i.e. X & W) were significantly higher than to the uncorrelated cues (i.e. Y & Z), $F(1,105)=20.42$, $p<.001$, $\eta_p^2=.163$. There was also a main effect of group, indicating a general tendency to judge the redundant cues as being more likely causes among preventative participants than among non-preventative participants, $F(1,105)=20.02$, $p<.001$, $\eta_p^2=.160$. Although the group x cue type interaction did not reach significance, $F(1,105)=1.68$, $p=.197$, $\eta_p^2=.016$, inspection of Figure 4.3e suggests a potentially greater redundancy effect following non-preventative pretraining than preventative pretraining. Subsequent paired samples t-tests of simple effects confirmed that the redundancy effect was significant in the non-preventative group, $t(56)=5.41$, $p<.001$, $d=.717$, but was non-significant in the preventative group, $t(49)=1.84$, $p=.072$, $d=.260$. This

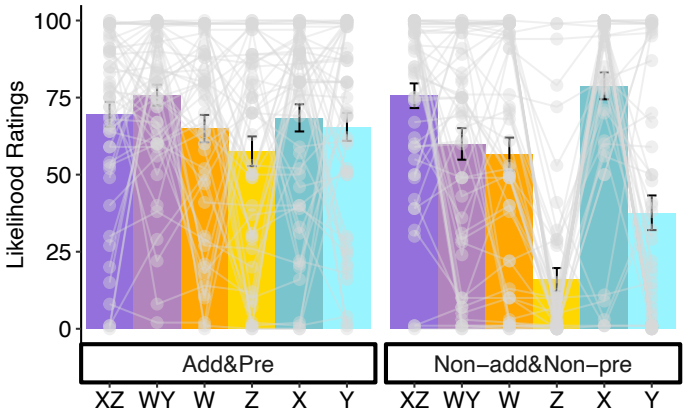
decrement in the magnitude of the redundancy effect only constituted weak evidence in favour of the null, according to the Bayesian binomial t-test, BF_{10} =.734.

Experiment 4.4 Confidence. The same ANOVA with confidence ratings as the dependent variable revealed a significant group x cue type interaction, $F(1,105)=5.46$, $p=.021$, $\eta_p^2=.049$. This suggests that the greater judgment certainty for the uncorrelated cues than for the blocked cues depended on whether preventative relationships were viewed as plausible. However, the main effect of group, $F(1,105)=2.81$, $p=.097$, $\eta_p^2=.026$, or cue type, $F(1,105)=3.46$, $p=.066$, $\eta_p^2=.032$, was neither significant. Further exploration of the interaction effect through paired samples t-tests indicated higher confidence for the uncorrelated cues than for the blocked cues in the non-preventative group, $t(56)=2.85$, $p=.006$, $d=.378$, and similar confidence for the critical redundant cues in the preventative group, $t(49)=.36$, $p=.719$, $d=.051$. Subsequent Bayesian binomial t-test confirmed that that the confidence difference was not present following preventative pretraining, $BF_{10}=.164$.

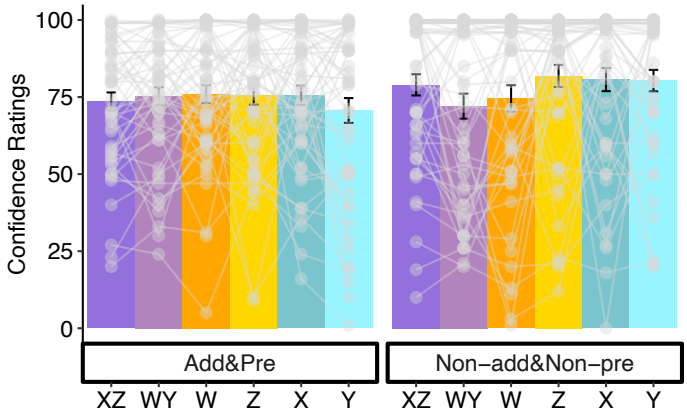
Figure 4.4

Mean likelihood ratings and confidence ratings for the critical cues and two test compounds on the Stage 2 ratings for each of Experiments 4.2-4.4. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

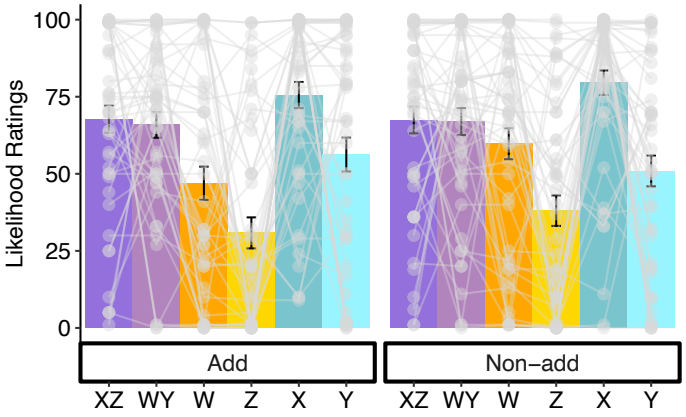
(a) Experiment 4.2 likelihood Rating



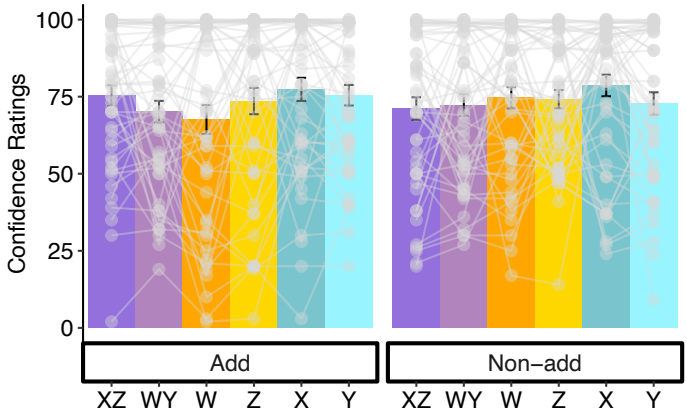
(b) Experiment 4.2 Confidence Rating



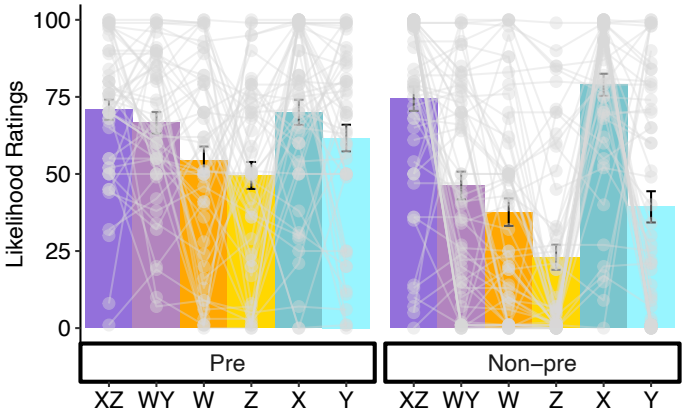
(c) Experiment 4.3 Likelihood Rating



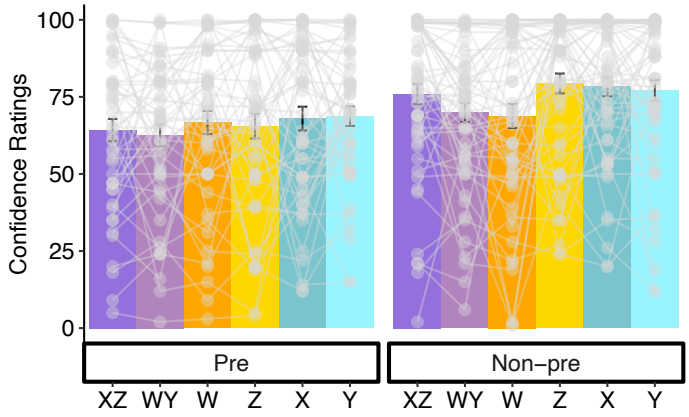
(d) Experiment 4.3 Confidence Rating



(e) Experiment 4.4 Likelihood Rating



(f) Experiment 4.4 Confidence Rating



Stage 2 Ratings Test

Likelihood and confidence ratings for the critical cues on the Stage 2 ratings test are illustrated in Figure 4.4. For each experiment (and consistent with Experiment 4.1), the comparisons of main interest were (i) between the critical compounds and (ii) between the individually presented redundant cues. The first comparison was analysed with a (2)x2 mixed model ANOVA with compound type as the within-subjects factor and pretraining condition as the between-subjects factor, while the second was analysed with a (2)x(2)x2 mixed model ANOVA with cue type and compound training as within-subjects factors and pretraining condition as the between-subjects factor. Results for each experiment are reported in sequence.

Experiment 4.2, Critical Compounds, Likelihood. The ANOVA results revealed a significant compound type x group interaction, $F(1,101)=9.17$, $p=.003$, $\eta_p^2=.083$, suggesting that the difference between XZ and WY (and hence, in learning about the blocked cue and the uncorrelated cue when trained as a compound) was influenced by pretraining. Neither the main effect of compound, $F(1,101)=1.68$, $p=.197$, $\eta_p^2=.016$, nor group, $F(1,101)=1.08$, $p=.302$, $\eta_p^2=.011$, was significant. Further t-tests showed that XZ rating was significant higher than WY (i.e. there was greater update of causal beliefs for X than for Y) in the non-additive and non-preventative group, $t(50)=2.84$, $p=.007$, $d=.397$. This difference did not reach significance in the additive and preventative group, $t(51)=1.33$, $p=.189$, $d=.185$, $BF_{10}=.347$.

Experiment 4.2, Critical Compounds, Confidence. There was a significant interaction between compound type and group, $F(1,101)=5.19$, $p=.025$, $\eta_p^2=.049$, suggesting that the differential impact of Stage 2 training on confidence for the blocked cue and the uncorrelated cue varied between pretraining groups. Neither the main effect of compound type, $F(1,101)=1.89$, $p=.172$, $\eta_p^2=.018$, nor group, $F(1,101)=.06$, $p=.807$, $\eta_p^2<.001$, was significant. Further t-tests indicated that confidence ratings were higher for XZ than WY

(certainty was improved more for X than for Y) as a consequence of Stage 2 learning for non-additive and non-preventative participants, $t(50)=2.20$, $p=.032$, $d=.308$, but not for the additive and preventative participants, $t(51)=.80$, $p=.427$, $d=.111$.

Experiment 4.2, Individual Cues, Likelihood. The blocked cues were on average judged as being more likely causes than the uncorrelated cues, $F(1,101)=46.91$, $p<.001$, $\eta_p^2=.317$. Additive and preventative participants gave higher ratings to cues in general than non-additive and non-preventative participants, $F(1,101)=16.09$, $p<.001$, $\eta_p^2=.137$. The significant cue type x group interaction, $F(1,101)=28.31$, $p<.001$, $\eta_p^2=.219$, suggests that the redundancy effect averaged over two critical cue pairs was stronger in the non-additive and non-preventative group than in the additive and preventative group. Likewise, the significant compound training x group interaction, $F(1,101)=8.75$, $p=.004$, $\eta_p^2=.080$, suggests that the higher average ratings to X and Y than to W and Z was more marked in the non-additive and non-preventative group than in the additive and preventative group. The cue type x compound training x group interaction was not significant, $F(1,101)=.19$, $p=.661$, $\eta_p^2=.002$. Further t-tests revealed that, among non-additive and non-preventative participants, the redundancy effect was present in both the critical pair that underwent compound training (X vs. Y), $t(50)=5.58$, $p<.001$, $d=.782$, and the pair that did not (W vs. Z), $t(50)=7.72$, $p<.001$, $d=1.081$, but among additive and preventative participants, there was no difference between X and Y, $t(51)=.50$, $p=.618$, $d=.070$, $BF_{10}=.170$, nor between W and Z, $t(51)=1.26$, $p=.214$, $d=.174$, $BF_{10}=.317$.

Experiment 4.2, Individual Cues, Confidence. Analyses of confidence ratings revealed no significant main effects of cue type, compound training, group, or interaction effects between these variables, with all *p*s>.08. Inspection of Figure 4.4b suggests that non-additive and non-preventative participants were more certain about the causal status of Z than for W, but this difference did not reach significance, $t(50)=1.76$, $p=.085$, $d=.246$.

Experiment 4.3, Critical Compounds, Likelihood and Confidence. Compounds XZ and WY were, on average, rated as being equally likely to cause the outcome, $F(1,95)=.09$, $p=.772$, $\eta_p^2<.001$, indicating no significant difference in learning between X and Y. This pattern was unaffected by whether pretraining was additive or non-additive, $F(1,95)=.03$, $p=.867$, $\eta_p^2<.001$. Likelihood ratings averaged over the two compounds were comparable between groups, $F(1,95)=.01$, $p=.929$, $\eta_p^2<.001$. Analyses of confidence ratings revealed no significant difference between XZ and WY, $F(1,95)=.79$, $p=.376$, $\eta_p^2=.008$. Although confidence ratings seem to exhibit an opposite trend for the additive group and the non-additive group, the interaction between cue type and pretraining did not reach significance, $F(1,95)=1.85$, $p=.177$, $\eta_p^2=.019$.

Experiment 4.3, Individual Cues, Likelihood and Confidence. Cues X and Y that received additional Stage 2 training were on average rated as more likely causes than W and Z, $F(1,95)=49.87$, $p<.001$, $\eta_p^2=.344$. The blocked cues X and W were regarded as more likely causes than the uncorrelated cues Y and Z averaged over compound training, $F(1,95)=36.80$, $p<.001$, $\eta_p^2=.279$. Although the average redundancy effect appears to be stronger for the XY pair than for the WZ pair, this difference was not statistically significant, $F(1,95)=3.03$, $p=.085$, $\eta_p^2=.031$. Confidence ratings to cues did not differ according to cue type, $F(1,95)=.13$, $p=.723$, $\eta_p^2=.001$, compound training, $F(1,95)=2.55$, $p=.114$, $\eta_p^2=.026$, or group, $F(1,95)=.13$, $p=.718$, $\eta_p^2=.001$. However, confidence for the blocked and uncorrelated cues differed significantly depending on whether the pair underwent compound learning, $F(1,95)=5.33$, $p=.023$, $\eta_p^2=.053$. Subsequent t-tests indicated that, averaged across groups, X was rated as a more confident cue than Y, $t(96)=2.02$, $p=.046$, $d=.205$, but W was given similar ratings as Z, $t(96)=1.17$, $p=.246$, $d=.118$.

Experiment 4.4, Critical Compounds, Likelihood. On average, there were higher ratings for XZ than WY (indicating a general learning advantage for X over Y) in both groups, $F(1,105)=18.74$, $p<.001$, $\eta_p^2=.151$. The main effect of group was also significant, $F(1,105)=4.26$ $p=.041$, $\eta_p^2=.039$, indicating greater overall likelihood ratings given by preventative participants than by non-preventative participants. Importantly, there was a significant compound type x group interaction, $F(1,105)=10.22$ $p=.002$, $\eta_p^2=.089$, suggesting that the learning bias toward X over Y was significantly more pronounced in the non-preventative group than in the preventative group. Subsequent paired samples t-tests revealed that ratings for XZ were higher than for WY in the non-preventative group, $t(56)=4.60$, $p<.001$, $d=.609$, but not in the preventative group, $t(49)=1.09$, $p=.280$, $d=.154$.

Experiment 4.4, Critical Compounds, Confidence. Confidence ratings for XZ and WY were, on average, higher for the non-preventative participants than preventative participants, $F(1,105)=4.89$ $p=.029$, $\eta_p^2=.044$. The main effect of compound type was not significant, $F(1,105)=3.17$ $p=.078$, $\eta_p^2=.029$. The interaction between group and compound type was not significant either, $F(1,105)=1.06$, $p=.305$, $\eta_p^2=.010$. However, inspection Figure 4.4f shows slightly higher confidence for XZ than for WY in the non-preventative group, which is not the case for the preventative group. Paired samples t-tests confirmed that non-preventative participants judged XZ with higher confidence than they did for WY, $t(56)=2.15$, $p=.036$, $d=.285$, which was not observed for preventative participants, $t(49)=.49$, $p=.626$, $d=.069$.

Experiment 4.4, Individual Cues, Likelihood. On average, participants gave higher likelihood ratings to the blocked cues than to the uncorrelated cues, $F(1,105)=34.03$, $p<.001$, $\eta_p^2=.245$, higher likelihood ratings to the additionally trained redundant cues X and Y than to cues W and Z, $F(1,105)=58.00$, $p<.001$, $\eta_p^2=.356$, as well as generally higher likelihood ratings in the preventative group than in the non-preventative group, $F(1,105)=13.99$,

p<.001, η_p^2=.118. The overall redundancy effect was significantly stronger in the non-preventative than in the preventative group, F(1,105)=12.62, p<.001, η_p^2=.107, and stronger between X and Y, the pair trained in Stage 2, than between W and Z, F(1,105)=6.94, p=.010, η_p^2=.062. The higher likelihood ratings to X and Y than to W and Z were significantly more pronounced in the non-preventative group than in the preventative group, F(1,105)=7.10, p=.009, η_p^2=.063. Moreover, a marginal three-way interaction suggested that the greater redundancy effect following compound training was significantly more marked among non-preventative participants than among preventative participants, F(1,105)=3.95, p=.050, η_p^2=.036.

Experiment 4.4, Individual Cues, Confidence. Non-preventative participants were, on average, more confident about the causal status of the redundant cues than preventative participants, F(1,105)=5.29, p=.023, η_p^2=.048. Neither the main effect of cue type, F(1,105)=2.00, p=.161, η_p^2=.019, nor compound training, F(1,105)=2.09, p=.151, η_p^2=.020, was significant. All interactions were non-significant, with all Ps>.09. However, Figure 4.4f shows that non-preventative participants gave slightly higher confidence ratings to Z than to W, which was not the case for preventative participants. Subsequent paired samples t-tests revealed higher certainty for Z over W in the non-preventative group, t(56)=2.55, p=.014, d=.337, but no difference was found in the preventative group, t(49)=.34, p=.734, d=.048.

Discussion

In summary, Experiments 4.2-4.4 investigated the effects of various pretraining conditions on the theory protection effect, using a similar but more balanced design compared to Experiment 4.1. Across these experiments, which used an identical training and test procedure (only differing in the pretraining and initial instructions), differences in the causal ratings for XZ and WY compounds at Stage 2 test provide the strongest test of theory protection. We examined how this difference varied with pretraining, and how it related to

ratings for blocked and uncorrelated cues after Stage 1, and individual cue ratings after Stage 2.

Consistent with Experiment 4.1, we found clear evidence of the theory protection effect in compound likelihood ratings (XZ>WY) after non-additive and non-preventative pretraining (Experiment 4.2), and after non-preventative pretraining (Experiment 4.4). This effect was not evident after additive and preventative pretraining (Experiment 4.2), preventative pretraining (Experiment 4.4), or in Experiment 4.3, where we manipulated additivity assumptions (note that in Experiment 4.3, preventative properties of the cues were possible though not emphasised using pretraining).

Despite the modifications to the design, made in an attempt to improve the chances of observing low-but-confident ratings for blocked cues, in no condition was there any evidence of a reversal of the effect (WY>XZ) that might be expected if participants were confidently concluding that the blocked cue was not causal. We hypothesised that this reversal would be most likely after Additive and Preventative pretraining (Experiment 4.2) and Additive pretraining (Experiment 4.3). In these conditions, ratings for the two compounds were equivalent. In Experiment 4.3, additive pretraining resulted in lower likelihood ratings for the blocked cues relative to non-additive pretraining, which is consistent with the hypothesis that deductive reasoning about the blocked cues was encouraged in the additive pretraining group. However, the size of the redundancy effect (i.e. X/W vs Y/Z) did not differ between groups and there was little other evidence across Experiments 4.2 and 4.3, in either the likelihood or confidence ratings, to suggest that additive pretraining had reduced causal beliefs and increased confidence about the blocked cues. We will reserve further speculation about this result for the General Discussion.

The theory protection effects observed under non-preventative pretraining in Experiments 4.2 and 4.4 were accompanied by significant increases in the magnitude of the

redundancy effect (larger differences between likelihood ratings for X/W vs. Y/Z) after Stage 1 training. This enhanced redundancy effect was also observed at Stage 2 in both experiments (manifesting in significant interactions between pretraining group and cue type). With the additional XY+ training in Stage 2, we might have expected the size of this redundancy effect to be even further enhanced for the X vs. Y comparison compared to W vs. Z. Evidence for this was weak (the three-way interaction was non-significant in Experiment 4.2, and marginal in Experiment 4.4). However, this comparison examines differences in ratings for cues that reside at different points on the response scale, and this weakness is the very reason for implementing the compound test procedure in the first place. The presence of the theory protection effect and enhanced redundancy effect in the non-preventative pretraining conditions was accompanied by *some* evidence of changes in confidence after Stage 1 training. In Experiment 4.4, non-preventative pretraining significantly enhanced confidence differences (Y/Z > W/X) relative to preventive pretraining. However, the evidence for this effect on confidence in Experiment 4.2 was much weaker, with the interaction not approaching significance. We will reflect more on the relevance of this inconsistent result in the General Discussion. Please note that K-means clustering analysis was not conducted for this chapter because participants did not exhibit a strong dissociative judgment pattern between causal likelihood and prediction certainty. Rather than relying on confident elimination, the subsequent bias in XY+ training appears to be largely driven by the propensity to protect existing theory that Y is non-causal (derived via elimination of preventative influences, but not always an obvious increase in confidence).

Table 4.3

Summary of key results relevant to theory protection

Expt	Condition	Stg 1 R.E. (X / W > Y / Z)	Stg 1 confidence (Y / Z > X / W)	Stg 2 T.P. (XZ > WY)
1	neutral	*d*=.352	*d*=-.071	*d*=.147
1	add. & prev.	*d*=.166	*d*=-.128	*d*=-.019
1	non-add. & non-prev.	*d*=1.191	*d*=.674	*d*=.636
2	add. & prev.	*d*=.167	*d*=-.029	*d*=-.185
2	non-add. & non-prev.	*d*=1.113	*d*=.214	*d*=.397
3	add.	*d*=.323	*d*=.211	*d*=.048
3	non-add.	*d*=.650	*d*=-.070	*d*=.012
4	prev.	*d*=.260	*d*=-.051	*d*=.154
4	non-prev.	*d*=.717	*d*=.378	*d*=.609

Note. Values represent effect sizes (Cohen's d) for simple comparisons in each group, using paired samples t tests. Values in light grey were non-significant.
Expt=Experiment number; R.E.=Redundancy effect, T. P.=Theory protection effect; add. = additive, non-add. = non-additive, prev. = preventative, non-prev. = non-preventative.

General discussion

The present study investigated the factors that drive theory protection in causal learning, that is, the tendency to learn more about one cue over another, based on learners' prior knowledge about each cue. Across four experiments, learners' causal assumptions were systematically manipulated in a pretraining phase accompanied by explicit instructions. We targeted assumptions that permit (or prohibit) the use of deductive reasoning about causally ambiguous cues. Via this pretraining, we sought to establish situations in which the participant could confidently judge a cue to be non-causal. When the two redundant cues subsequently entered into Stage 2 training as a compound, the cue whose causal ambiguity

had been resolved through deductive inferences formed in Stage 1 training should undergo less learning than the other cue whose causal status remained ambiguous. Table 4.3 shows a summary of the key results from the current chapter.

Experiment 4.1 demonstrated a greater update of existing causal knowledge about a blocked cue than about an uncorrelated cue, replicating Spicer et al. (2020). However, this pattern was only found when pretraining emphasized non-additive and non-preventative beliefs, which should encourage deductive reasoning about the uncorrelated cue and limit deductive reasoning about the blocked cue. This was the only condition under which confidence was higher for the uncorrelated cue than for the blocked cue and the only condition in which we observed differential learning consistent with the theory protection principle. The certainty difference between the redundant cues was not observed when pretraining highlighted additive and preventative assumptions or was neutral in assumptions, which appeared to abolish any learning difference in the subsequent compound training.

Experiment 4.2 replicated the blocked cue learning bias both in the form of the redundancy effect in Stage 1 training and the enduring learning advantage in Stage 2 compound training. Again, these effects were observed specifically among non-additive and non-preventative participants, albeit with a lack of clear inverse relationship between likelihood and confidence judgments in Stage 1. These results suggest that causal certainty may influence compound learning under certain circumstances, but its influence must be secondary to likelihood judgment which was more consistently modulated by deductive reasoning. The effect of assumptions targeting the blocked and uncorrelated cues was disentangled in Experiments 4.3 and 4.4. Although additive pretraining resulted in the blocked cues being regarded as a less likely cause than after non-additive pretraining, the modulating effect of magnitude additivity assumptions did not have a strong impact on the strength of the redundancy effect or the extent of subsequent learning in the blocked-

uncorrelated cue (XY) compound. In contrast, Experiment 4.4 found a much stronger influence of pretraining focusing on the preventative vs. non-preventative nature of the outcome. Non-preventative pretraining led to the uncorrelated cue being regarded as a much less likely cause and associated with much higher confidence than the blocked cue. This group also displayed a much stronger Stage 2 learning bias toward learning about the blocked cue.

On the whole, the current findings suggest that the dissociative pattern elicited by the redundancy effect, that is, higher likelihood judgment for the blocked cue than for the uncorrelated cue but lower confidence judgment for the blocked cue than for the uncorrelated cue, may not be necessary to produce the Stage 2 learning bias. However, a large enough likelihood judgment difference seems to be the key in the biased attribution of outcome in compound training.

The theory protection hypothesis proposed by Spicer et al. (2020) is that when people experience the XY+ compound, they preferentially attribute causation to X if (and only if) they possess a strong belief that Y is non-causal but are unsure whether X is causal or not. Taken together, the current findings are not fully consistent with this view. Instead, the current findings imply a tendency to preserve existing non-causal relationships in situations when subsequent observations conflict with previously acquired knowledge. In the non-additive and non-preventative group of Experiment 4.1, the greater ease to apply deductive reasoning to ascertain the causal status of Y than for X resembles the typical situation in the standard redundancy effect particularly under the food allergist paradigm (e.g. Jones et al., 2019; Spicer et al., 2020; Uengoer et al., 2013) where there is a natural propensity to deduce the non-causality of Y but a reluctance to engage in reasoning for X. In keeping with Spicer and colleagues (2020), the unexpected occurrence of the outcome on subsequent XY+ trials was attributed more to the causally ambiguous X as a means to preserve the confidently

established association between Y and outcome non-occurrence. Although there was weak evidence of higher likelihood judgment for X than for Y in the neutral group where assumptions were not manipulated in pretraining, certainty in likelihood judgments did not differ between the critical cues. The lack of preceding confidence difference seems to have abolished any learning difference during compound training that would otherwise be expected to follow the observation of the redundancy effect. However, participants given prior training on non-additive and non-preventative assumptions in Experiment 4.2 demonstrated a greater update in causal knowledge for X than for Y despite the fact that the redundancy effect was observed in the absence of differing confidence. Although it is unclear as to why the same pretraining manipulation more effectively enhanced deductive inferences among non-additive and non-preventative participants in Experiment 4.1 than in Experiment 4.2, evidence of unequal learning about the blocked and uncorrelated cues both with and without prior confidence differences suggests that causal uncertainty cannot be a necessary driving factor for the amount of updating that the redundant cues undergo in compound training.

Alternatively, likelihood judgment and prediction certainty may be independent consequences of deductive reasoning that are not closely linked to one another. Although the presence of the typical dissociative pattern would convincingly suggest the involvement of inferential processes in the redundancy effect, it may not be necessary that the causal beliefs formed on the basis prior propositions are held with stronger confidence than that derived from associative principles. The inference that the uncorrelated cue is non-causal on its own may lead to the theory protection result in the Stage 2 compound training regardless of how strongly the inference is formed. Thus, the tendency to protect existing causal knowledge may be reliant any kind of deductive inference that reduces causal significance of the uncorrelated cue and increases the conflict between the uncorrelated cue and the Stage 2

outcome. However, it must be conceded that the confidence measure used in the current experiments may not be sensitive enough to detect subtle differences in causal ambiguity and the lack of consistent confidence difference between redundant cues may be a result of type 2 error. Future studies using a better confidence test would allow for a more scrupulous assessment of causal ambiguity.

Evaluation of the independent contribution that magnitude additivity assumptions and preventative cue assumptions make to compound learning provides further evidence that prediction error (i.e. between the outcome predicted by the cue and what actually happened) is responsible for theory protection. Manipulating assumptions for the blocked cue in Experiment 4.3 resulted in lower likelihood judgment for the blocked cue among additive participants than among non-additive participants, but did not explicitly enhance the belief that the uncorrelated cue is non-causal. As a result, Y evoked a similar degree of inconsistency when confronted with the outcome in compound training among both groups of participants. There was no evidence of a learning bias favoring either X or Y in Stage 2 in either of these groups. By contrast, adjusting assumptions for the uncorrelated cue in Experiment 4.4 reduced likelihood judgment but enhanced confidence for the target cue when preventative cue-outcome relationships were deemed implausible. The strong belief that Y cannot be causal following non-preventative pretraining elicited a sufficiently large prediction error on compound training trials that acted to bias learning toward X.

Note that this type of error-driven learning is different from individual prediction error accounts which expect Y to acquire a larger gain in associative strength compared to X (Rescorla, 2001). It may be argued that, in line with theory protection, the improved certainty for Y as a secondary product of deductive reasoning is capable of biasing learning toward X in an attempt to preserve the established relationship between Y and outcome non-occurrence in this instance. However, results from Experiment 4.2 suggests that an enhanced likelihood

judgment difference generated by the deduction that the uncorrelated cues are non-causal is sufficient for the blocked cue learning advantage in compound training.

An unusual finding that was observed in Experiment 4.1 was the anomalously high likelihood ratings to the blocked cue following additive and preventative pretraining. This peculiarity points to an issue of overemphasis on preventative assumptions which may have inadvertently made a different type of deduction possible for the blocked cues: knowledge of A and D as individual causes and the fact that their effect remained valid with the addition of X and W to them allowed the inference that X and W must be causes. That is, if X and W did not prevent the effect of co-present others, they must be causal. Experiment 4.2 sought to weaken the unintended emphasis on preventative relationships by demonstrating its unstable nature such that prevention occurred on some of the trials but not on others. Additive pretraining in this experiment reduced likelihood judgment for the blocked cues but did not weaken the redundancy effect, suggesting possible influence of additive assumptions on the uncorrelated cues (i.e. Y cannot be causal because it did not lead to a larger effect together with C). These are noteworthy points for future investigations of magnitude additivity assumptions to consider.

Alternatively, it is possible that certainty with which the redundant cues are judged influences the amount of associative update in compound learning in a way that is different to the one strictly specified by theory protection. Consider the three possible cases where X has been rated as more likely to cause an outcome than is Y (in line with the redundancy effect): First, the learner is less confident about X than Y; second, the learner is *more* confident about X than Y; and third, the learner is equally confident about X and Y. The first case is the ideal situation to trigger theory protection in which existing knowledge about Y being a non-cause is protected by assigning the likely causal agent to the ambiguous X during XY+ compound training. In the second case, observing that XY is followed by the outcome would not

necessarily contradict beliefs about X; if the learner is confident that X is more likely to cause the outcome than Y then XY+ trials may serve to strengthen the belief that X is causal, while leaving Y causally ambiguous. Although the third case appears to imply a similar amount of learning about the equally certain X and Y, the presence of the outcome is nonetheless more consistent with the prediction elicited by X based on previously acquired causal knowledge (i.e. Y better anticipates outcome absence). The greater predictive utility of X with respect to outcome presence then leads to greater learning about X. It is clear from the above analysis that confidence difference in any direction *could* foster a learning advantage for X over Y so long as Stage 1 training led the learner to believe that X is more likely to result in the outcome than is Y.

As alluded to in the forgoing analysis, the reliably demonstrated redundancy effect across all conditions that highlighted non-preventative assumptions suggests possible explanations based on mere judgment difference between the redundant cues. If this were the case, then the compound test result should always yield greater learning about X over Y as long as the blocked cue was judged to be a more likely cause than the uncorrelated cue. However, as observed in Experiments 4.1 and 4.3, this result was *not* obtained when the redundancy effect was elicited without enforcing an assumption in neutral pretraining or by imposing magnitude additivity assumptions. These conditions all produced relatively modest redundancy effects compared to the differences achieved through non-preventative pretraining, and it could be the case that a large likelihood judgment difference between the blocked and uncorrelated cues is the key to differential learning in compound training. Future studies are encouraged to implement an additive and non-preventative pretraining phase that reduce likelihood judgment for both redundant cues, which would offer a more direct test for whether the observation of the redundancy effect is critical in addition to low likelihood judgment for the uncorrelated cue. Likewise, a more refined procedure to modify

assumptions specifically targeting the blocked cue may provide further insight into whether an amplified redundancy effect achieved *without* lowering likelihood judgment for the uncorrelated cue would suffice for the blocked cue learning advantage in compound training.

The ease with which the non-causality of the uncorrelated cue is inferred from non-preventative assumptions is a recurring theme that dominates interpretations of the current results. Under the hormone change paradigm, medicines are generally considered capable of preventing the effects of other medicines taken at the same time, and as such, in the absence of explicit emphasis on non-preventative relationships, the uncorrelated cues will be regarded as moderate causes whose effects may be obstructed by B on BY trials. The food allergist task, on the other hand, brings with it the widely held belief that the effect of an allergenic food cannot be counteracted by consuming other foods together. The readiness to deduce the non-causal role of the uncorrelated cue in the food allergist task thus aligns with conditions that encourage non-preventative assumptions in the hormone change task. It seems reasonable to speculate that Spicer et al. (2020) found greater learning about the blocked cue in compound training without experimentally manipulating assumptions for the uncorrelated cue because the inference that the uncorrelated cue is non-causal is naturally enabled by the food allergist scenario (Jones et al., 2019; Zaksaite & Jones, 2020). Future studies could compare the effect of assumption manipulations in different task paradigms.

While it is a core assumption underlying ambiguity- and error-driven learning that associative strengths on XY+ trials are unequally distributed between X and Y, other theorists have argued that learning follows common error learning rules (Rescorla &Wagner, 1972) but additionally assume a type of disproportionate mapping of associative change to consequence on performance (Holmes et al., 2019; Chan et al., 2021). Specifically, these authors devised a double sigmoid function that maps the same amount of change in associative strength onto varying contributions to performance depending on a cue's initial

associative state, thereby obviate the need for varying uncertainty or predictiveness between the redundant cues. The redundancy effect reported by Spicer et al. (2020), and other human causal learning studies under a food allergist paradigm represents a typical case of judgment difference that involves a moderately likely cause X and a very unlikely cause Y. The intermediate rating to X would locate it at the increasing part of the function, while the low rating to Y would locate it at the decreasing part of the function. The same increment in associative strength for X and Y during XY+ training would therefore translate into a larger contribution to performance for X than for Y. The current study replicated the typical middling rating to X and low rating to Y following explicit encouragement of non-preventative assumptions, and in keeping with Holmes et al. (2019), these participants subsequently demonstrated a larger performance increment for X than for Y on test.

The double sigmoid mapping function fares less well in atypical situations of the redundancy effect. For example, under the hormone change paradigm, the judgement difference was evoked by a moderate cause X and a weak cause Y for participants given neutral pretraining in Experiment 4.1, and was evoked by a strong cause X and a weak cause Y for participants given non-additive pretraining in Experiment 4.3. In these two conditions, the causal X would be placed at the decreasing part of the curve, while the slightly causal Y would be placed at the increasing part of the curve. The same associative increment for X and Y in compound training should thus result in a larger improvement in performance for Y than for X. This prediction is inconsistent with the observed equivalent impact on performance for X and Y in the neutral group and the non-additive group.

The partial consistency of the current findings observed with the theory protection principle also holds true for attentional models of learning. In particularly, the differential associative change for X and Y may be explained by assigning a role to the amount of attention afforded to each redundant cue during compound training (Mackintosh 1975;

Kruschke, 2005). In Experiments 4.1, 4.2, and 4.4, attention may be preferentially devoted to X because it possesses higher predictive utility than Y according to Mackintosh's model of selective attention or be strategically reallocated to X in an attempt to minimise the large prediction error generated by Y according to Kruschke's EXIT model. In either case, learning about the poorer predictor Y will be overshadowed by the better predictor X, leading to a greater associative increment for X. However, attentional accounts do not explain why the blocked cue did not undergo greater learning in compound when the redundancy effect was elicited without deliberately reducing causal significance of the uncorrelated cue. More generally, accounts based solely on prediction error learning (with or without attention modulation) do not provide a satisfactory explanation for why pretraining influences learning about X and Y in the first place.

In conclusion, the reported findings coherently support a crucial role for prior likelihood judgment about the uncorrelated cue in biased learning about these cues when presented in compound. Dissociative likelihood judgement and causal certainty, proposed as an essential prerequisite for theory protection, was not necessary to drive the subsequent blocked cue learning benefit. Instead, encouraging non-preventative assumptions was both necessary and sufficient to produce biased learning toward the blocked cue in Stage 2. These results suggest that compound learning of redundant cues is not critically dependent on the blocked cue being more causally ambiguous than the uncorrelated cue, but rather, on whether the hallmark judgment difference itself is brought about by judgement of the uncorrelated cue as an unlikely cause.

Chapter 5

The Choice of Cover Story

It has been a recurring theme throughout the current thesis that the choice of task scenario may influence judgment of redundant cues. The conventional food allergist task requires participants to predict allergic reactions from certain combinations of foods eaten. This task has been commonly used across cue competition paradigms such as blocking and overshadowing (e.g. Lovibond et al., 2003; Mitchell et al., 2006), the learned predictiveness effect (e.g. Le Pelley & McLaren, 2003), and the inverse base-rate effect (e.g. Don et al., 2019). The widespread usage has led researchers to adopt the same cover story for the redundancy effect upon its initial discovery in human causal learning (Uengoer et al., 2013). In more recent years, however, the suitability of the food allergist task began to be questioned by studies of non-causal relationships (Jones et al., 2019; Zaksaite & Jones, 2020). These later authors found that human participants tend to be particularly confident about non-causality, deeming cues to be non-causal even if they were only occasionally paired with outcome absence.

The elevated confidence about non-causal relationships reflects a failure to entertain (or decision to discount) the possibility that accompanying cues may prevent an outcome from occurring. This is of particular relevance to learning about the uncorrelated cue in a relative validity design (Y in BY–/CY+). Regarding the food allergist task, cases of allergies being prevented by the consumption of another food are virtually unheard of in real life, as a result, learners appear to disregard any potential preventative influence from foods consumed together with Y (i.e. B) and take the absence of the outcome on BY trials as strong evidence of both B and Y failing to cause the ailment. This line of reasoning would not affect judgments of the blocked cue (X in A+/AX+) because X is always followed by the outcome. Even in conditions where AX is only probabilistically followed by the outcome, it is unlikely

a participant would discount X in the same way as Y because the participant never has the opportunity to observe the effect of X without A (in contrast, the participant has the opportunity to observe that Y has no effect in the absence of C). Therefore, the food allergist task may generate low likelihood judgement about Y and hence a large redundancy effect by virtue of the non-preventative assumptions inherent in the task itself.

The hormone change task was designed to tackle this issue by making preventative relationships more plausible (Zaksaite & Jones, 2020). That is, it is more natural for medicines to prevent hormone change than it is for foods to prevent allergic reactions. Given that learning about Y is particularly sensitive to preventative relationships, the current thesis opted for a causal context in which preventative relationships were explicitly plausible. This notable difference in the cover story might have contributed to the unexpected absence or reversal of the relative validity effect and the redundancy effect. Many previous demonstrations of the relative validity effect (Wasserman, 1990; Baetu et al., 2005; Callejas-Aguilera & Rosas, 2010) and the redundancy effect (Jones & Zaksaite, 2018; Uengoer et al., 2013, 2019, 2020) have used an allergist task in which the possibility that a food might prevent an allergy is unlikely to be entertained. Since B would be more likely to be interpreted as preventing the effect of Y in the hormone change task, participants might have given higher causal ratings to Y, effectively abolishing the relative validity effect and the redundancy effect.

Different task scenarios give varying degrees of emphasis on non-preventative relationships. While the food allergist task by its inherent nature validates the deductive inference that cues followed by outcome absence are non-causes, the hormone change task imposes additional restrictions on this simple deduction by making preventative cue influences plausible. Judgement about the partially reinforced uncorrelated cue was expected to be particularly susceptible to this difference between task scenarios, demonstrating as

lower likelihood judgment and higher confidence judgement in tasks that discourage prevention. In this study, we sought to confirm our hypothesis that people would reason about the uncorrelated cue differently depending on the causal scenario. We sought to show this with contingencies presented in written summary form to ensure that any differences between scenarios were not due to inherent differences in contingency learning (e.g. caused by differences in the associability of the cues and outcomes) but rather due to the inferences drawn about the information.

The involvement of propositional reasoning may be further supported by the demonstration of parallel between actual contingencies experienced through training trials and contingency information presented as a written summary. Previous studies have shown that contingency knowledge delivered in the form of written instructions can lead to the blocking effect (Lovibond, 2003) and the inverse base-rate effect (Johansen et al., 2007; but see Don et al., 2021) that are essentially equivalent to the experience of multiple training trials.

In this series of two experiments, judgments made under the conventionally used food allergist scenario and the relatively novel hormone change scenario. The judgment of most interest was for the uncorrelated cue from a relative validity design (as described in summary form). We examined the effect of causal scenario using summary descriptions of the redundancy effect (Experiment 5.1) and the relative validity effect (Experiment 5.2). Since the experiments are identical in all respects except the summary information provided, the methods for both are described together below.

Experiments 5.1 and 5.2

Method

Participants

First-year psychology students from the University of Sydney were recruited to participate in both experiments. One hundred and ten were assigned to Experiment 5.1, with 55 (47 females, mean age=18.87, *SD*=1.26) in the food allergist task first group and another 55 (41 females, mean age=19.05, *SD*=2.22) in the hormone change task first group. One hundred and thirteen were assigned to Experiment 5.2, with 56 (42 females, mean age=19.25, *SD*=2.59) in the allergist task first group and 57 (40 females, mean age=19.11, *SD*=2.27) in the hormone change task first group.

Design and Procedure

Contingencies that give rise to the redundancy effect and the relative validity effect were organised in summary descriptions (see supplementary materials for full details). In each of Experiments 5.1 and 5.2, participants completed the food allergist task first and the hormone change task second or vice versa to control for order effects. Participants in Experiment 5.1 received descriptions for trials of the kind, A+, AX+, BY–, CY+, K–, and KL–, representing the contingencies of a simple redundancy effect design (with fillers to balance the probability of the outcome). Participants in Experiment 5.2 received descriptions for trials of the kind BY–, CY+, FG+/–, and HG+/–, representing a simple relative validity design.

For the food allergist scenario, participants were asked to imagine that they were an allergist trying to determine the cause of an allergic reaction shortly after their patient ate meal. For the hormone change scenario, participants were instructed to imagine that they were a medical researcher trying to determine the effect of various medicines on a patient's hormone levels. Information about individual trial types were provided in text form,

summarising 10 observations. Rather than using fictitious medicine and real food names, the cues were presented as "Medicine/Food A". For instance, A+ trials might appear in the hormone change scenario as "On 10 separate days, the patient took Medicine A and each time experienced an INCREASE in hormone levels.", whereas FG+/– trials in the food allergist scenario would be expressed as "On 10 separate days, the patient ate Food G and Food F, and experienced an ALLERGIC REACTION on HALF of the days." (full details are provided in supplementary materials). The two cues of primary interest (X and Y in the redundancy effect; Y and G in the relative validity effect) were always labelled medicine/food "B" and "D" respectively.

After reading the contingency information, participants proceeded to a ratings test stage where likelihood and confidence ratings were recorded in keeping with the previous experiments in Chapters 3 and 4 (except that all measurements were taken once only in these experiments). The summary information for all contingencies was available on screen while participants made their ratings.

Results

For each experiment, the judgments of key relevance were analysed for their sensitivity to causal context. For each effect, likelihood and confidence judgements about critical redundant cues were assessed with a separate 2x(2x2) mixed model ANOVA with scenario order (food allergist task first vs. hormone change task first) as the between-subjects factor, and task scenario (food allergist task vs. hormone change task) and cue type (X vs. Y for the redundancy effect in Experiment 5.1; Y vs. G for the relative validity effect in Experiment 5.2) as within-subjects factors. Likelihood ratings given to individual cues at test were the dependent variable that indexed the strength of causal relationship extracted from contingency information. Figure 5.1 illustrates likelihood and confidence ratings for the redundancy effect and Figure 5.2 illustrates likelihood and confidence ratings for the relative

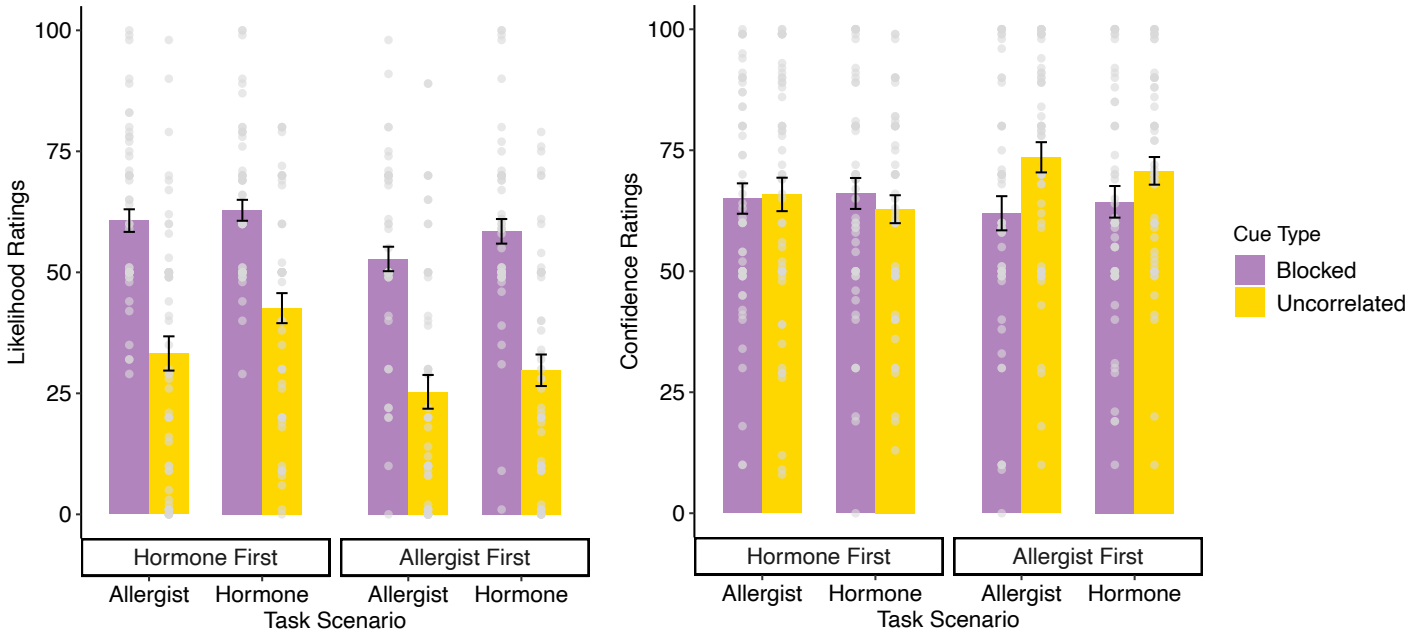
validity effect. Please see supplementary materials for complete figures including all cues for Experiments 5.1 and 5.2.

Experiment 5.1: The Redundancy Effect

Figure 5.1

(a) Mean likelihood ratings and (b) Mean confidence ratings for the redundancy effect tested in the food allergist task and the hormone change task with counterbalanced orders in Experiment 5.1. The middle two pairs of bars illustrate the first set of ratings made by participants (i.e. medicine ratings for the hormone change task first group and food ratings for the food allergist task first group). Dots indicate scores from individual participants and error bars indicate standard error of mean (SEM).

(a) (b)



Likelihood Ratings. There was a main effect of scenario, suggesting that the food allergist cover story produced a general tendency for participants to rate critical cues as less likely causes compared to the hormone change task, $F(1,108)=97.92$, $p<.001$, $\eta_p^2=.476$. The main effect of cue was significant, indicating an overall learning bias toward the blocked cue over the uncorrelated cue (i.e. X>Y), $F(1,108)=474.53$, $p<.001$, $\eta_p^2=.815$. The main effect of scenario order was also significant, with lower ratings on average in the group of participants who rated the allergist scenario first, $F(1,108)=7.44$, $p=.007$, $\eta_p^2=.064$. Importantly, scenario x cue interaction revealed a significantly stronger redundancy effect in the food allergist task than in the hormone change task regardless of test order, $F(1,108)=159.17$, $p<.001$, $\eta_p^2=.596$. Although the redundancy effect was numerically weaker in the first scenario than in the second, the scenario order x cue interaction did not reach statistical significance, $F(1,108)=3.68$, $p=.058$, $\eta_p^2=.033$. Neither the scenario x scenario order interaction, $F(1,108)=2.41$, $p=.123$, $\eta_p^2=.022$, nor the scenario x cue x scenario order interaction, $F(1,108)=.02$, $p=.883$, $\eta_p^2<.001$, was significant. Subsequent analyses of simple main effects confirmed that the redundancy effect was significant in both the food allergist task, $t(109)=9.83$, $p<.001$, $d=.937$, and the hormone change task, $t(109)=9.27$, $p<.001$, $d=.884$, and when tested first, $t(109)=8.43$, $p<.001$, $d=.804$, and second, $t(109)=10.83$, $p<.001$, $d=1.033$.

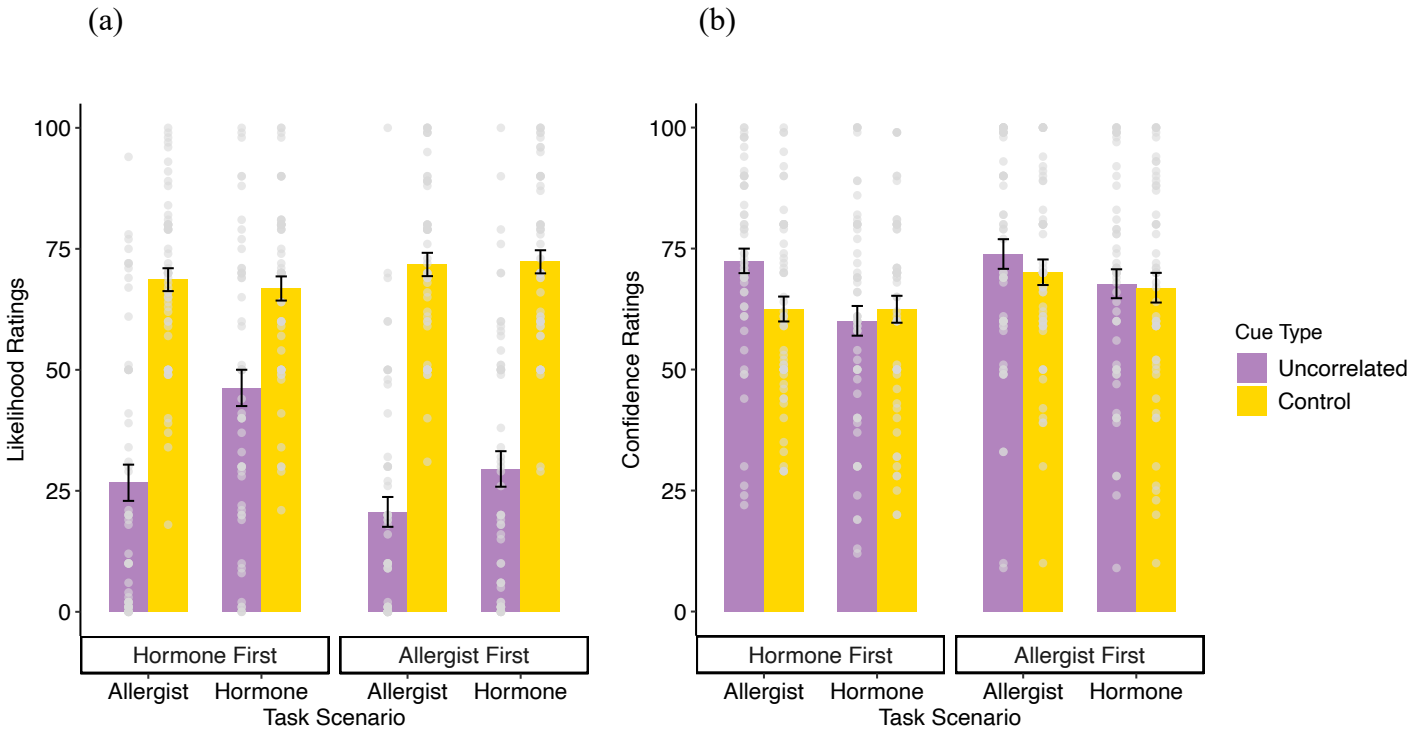
Confidence Ratings. Although the main effect of cue was non-significant, $F(1,108)=3.32$, $p=.071$, $\eta_p^2=.030$, the cue x scenario order interaction indicated that the higher confidence ratings for the uncorrelated cue than for the blocked cue depended on the order in which the tasks were completed, $F(1,108)=5.69$, $p=.019$, $\eta_p^2=.050$. The confidence difference appears to also depend on the task scenario that the critical cues were trained under, with elevated confidence for the uncorrelated cue relative to the blocked cue being numerically stronger under the food allergist task than under the hormone change task. However, the cue x scenario interaction did not reach significance, $F(1,108)=2.46$, $p=.120$, $\eta_p^2=.022$. All other

main effects or interaction effects were non-significant, all other *p*s>.4. Simple t-tests revealed that confidence was only higher for the uncorrelated cue than for the blocked cue in the food allergist task, $t(109)=2.20$, $p=.030$, $d=.210$, but not in the hormone change task $t(109)=.65$, $p=.517$, $d=.062$. No difference in confidence ratings was revealed combining tasks that were completed first, $t(109)=1.54$, $p=.127$, $d=.147$, or second, $t(109)=1.41$, $p=.162$, $d=.134$.

Experiment 5.2: The Relative Validity Effect

Figure 5.2

(a) Mean likelihood ratings and (b) Mean confidence ratings for the relative validity effect tested in the food allergist task and the hormone change task with counterbalanced orders in Experiment 5.2. Dots indicate scores from individual participants and error bars indicate standard error of mean (SEM).



Likelihood Ratings. There was a main effect of scenario, with critical cues receiving higher likelihood ratings overall in the hormone change task compared to the food allergist task, $F(1,111)=15.82$, $p<.001$, $\eta_p^2=.125$. The uncorrelated cue was rated as being a less likely cause than its control cue across scenarios (i.e. Y<G), indicating the presence of an overall relative validity effect, $F(1,111)=230.87$, $p<.001$, $\eta_p^2=.675$. The main effect of scenario order was non-significant, $F(1,111)=1.97$, $p=.163$, $\eta_p^2=.017$. Of most interest, the relative validity effect was significantly more marked if embedded within the food allergist compared to the hormone change task, $F(1,111)=20.13$, $p<.001$, $\eta_p^2=.153$. Scenario order also interacted with ratings for critical cues with relative validity being stronger in the group that completed the allergist scenario first, $F(1,111)=9.31$, $p=.003$, $\eta_p^2=.077$. The scenario x scenario order interaction was non-significant, $F(1,111)=1.49$, $p=.225$, $\eta_p^2=.013$. The effect of the three-way interaction was marginal, suggesting that the greater relative validity effect in the food allergist task than in the hormone change task was slightly more pronounced when these scenarios were presented as the first task than if they were completed second, $F(1,111)=3.90$, $p=.051$, $\eta_p^2=.034$. Further analyses of simple effects confirmed that the relative validity effect was significant in both the food allergist task, $t(112)=16.07$, $p<.001$, $d=1.512$, and the hormone change task, $t(112)=9.31$, $p<.001$, $d=.876$, and both when tested first, $t(112)=10.60$, $p<.001$, $d=.997$, and second, $t(112)=13.89$, $p<.001$, $d=1.306$.

Confidence Ratings. The main effect of scenario was significant, suggesting that the food allergist scenario elevated participants' overall confidence compared to the hormone change scenario, $F(1,111)=13.35$, $p<.001$, $\eta_p^2=.107$. Although participants were more confident about their likelihood judgments for the uncorrelated cue than for the control, the main effect of cue was non-significant, $F(1,111)=2.45$, $p=.120$, $\eta_p^2=.022$. The cue x scenario interaction revealed that confidence for the uncorrelated cue might be higher than that for the control but it depended on the task scenario, $F(1,111)=8.34$, $p=.005$, $\eta_p^2=.070$. All other

effects were non-significant, all other *p*s>.07. Subsequent t-tests showed that participants only felt more confident about the uncorrelated cue than the blocked cue in the food allergist scenario, $t(112)=2.89$, $p=.005$, $d=.271$, or in the second task, $t(112)=2.21$, $p=.029$, $d=.207$, but not in the hormone change scenario, $t(112)=.345$, $p=.731$, $d=.032$, or in the first task, $t(112)=.292$, $p=.770$, $d=.028$.

General Discussion

The choice of cover story was hypothesised to have a particular bearing on inferences made about the uncorrelated cue. The redundancy effect and the relative validity effect, which critically rely on judgment about the uncorrelated cue, were evaluated for their sensitivity to the food allergist scenario and the hormone change scenario. Presenting contingency information in summary format was sufficient to elicit both effects. The stronger relative validity effect in summary form suggests that learners may be more selective in learning or remembering the trial types when they are experienced sequentially over time. For example, contingency information about B and C is retained better than Y in sequential learning but is remembered to a similar extent in summary format. Compared to the hormone change task, the food allergist task had a much greater impact in reducing likelihood judgment about the uncorrelated cue than it did for the blocked cue. As a result, both effects were larger under the food allergist scenario than under the hormone change scenario. Although these effects were evident in both task scenarios, only the food allergist task produced confidence patterns for the uncorrelated cue that clearly dissociated from its likelihood judgements. Consistent with the proposed role of the food allergist task in fostering deductive inference about the uncorrelated cue being non-causal, these findings suggest a strong propensity for learners to engage in inferential reasoning processes for the uncorrelated cue when preventative influences are naturally ruled out by the cover story. The extent to which the task scenario by its very nature allows for deductively valid inferences

can therefore create judgment differences between redundant cues for reasons that are not a consequence of contingency learning, *per se*, and are arguably outside the scope of standard models of associative learning. The relatively high likelihood judgement about the uncorrelated cue brought about by preventative assumptions implied in the hormone change task might explain some of the inconsistencies between the current experiments and previous demonstrations of the redundancy using the food allergist task (e.g. Uengoer et al., 2013) and previous demonstrations of the relative validity effect in non-human animals (e.g. Murphy et al., 2001a; Murphy et al., 2001b).

The fact that people possess prevailing assumptions about cause and effect when they come into an experiment does not automatically mean those assumptions cannot be challenged or ignored in the process of learning. Learners clearly *can* learn about preventative relationships between foods and allergic reaction outcomes when the learning task demands it. For instance, food cues have been trained successfully as conditioned inhibitors and these inhibitors appear to have similar properties to inhibitory cues trained under other causal scenarios that do not so readily rule out prevention (e.g. see Karazinov & Boakes, 2007, cf. Lee & Livesey, 2012). The food allergist task has also been used successfully to train more complex inhibitory relationships such as those found in negative patterning and biconditional discriminations (Shanks & Darby, 1998; Don et al., 2020; Livesey et al., 2019). Importantly, however, the relative validity task is an example where the prevailing non-preventative assumption is not *directly* contradicted by the cue-outcome contingencies, in fact arguably holding this assumption helps to disambiguate the causal role of cue Y and thus reduces the difficulty of the learner's decisions. It is likely, therefore, to play an important role in producing the pattern of causal and confidence ratings observed in previous studies.

Chapter 6

General Discussion

The present thesis interrogated human causal learning in ambiguous situations, in particular those involved in the *redundancy effect*. The effect is deemed theoretically challenging for a broad range of existing theories as none suffices as a satisfactory explanation for the findings reported in the literature thus far (Uengoer et al., 2013; Uengoer et al., 2020). It was the overarching theme of this research to understand the processes by which the characterising judgment pattern of the redundancy effect – that is, higher causal ratings for the blocked cue from a blocking design compared to the uncorrelated cue from a relative validity design – is achieved in human learners. This objective was pursued in two ways. First, mechanistic models developed within the framework of traditional associative analysis were simulated and compared for their plausibility as formal mathematical accounts of the redundancy effect. Second, inferential reasoning processes that might be involved in judging ambiguous causes were examined by manipulating assumptions critical for deductive inferences prior to learning about redundant cues. By assessing likelihood judgment about ambiguous cues, it was assumed that the strength of causal relationship between independent events are reflected in the perceived likelihood that a cue is an effective cause. This was complemented by self-reported confidence to assess the dissociative pattern uniquely predicted by the inferential approach. The conclusions drawn from the present work should, however, not be viewed in terms of confirming or disconfirming either of these approaches, but be viewed as an evaluation of their contribution in relation to human learning about ambiguous causes. Indeed, the overall findings can be effectively interpreted by assuming the involvement of both: while basic associative principles lay the foundation for learning, complex propositional reasoning shows a more powerful and enduring influence on causal judgment about redundant cues. In this final chapter, I will first summarise the key findings

from each chapter before considering several issues that are broadly relevant to each series of experiments.

Under the probabilistic conditions used in Chapter 2, differential learning about redundant cues is best explained by taking a common element approach. Uengoer et al. (2013) observed the redundancy effect using a design where both the blocked and uncorrelated cues were paired with the outcome with a 50% probability. This methodologically complex, yet theoretically interesting, finding motivated a series of experiments in Chapter 2 with outcome probability and base rate manipulations. Contrary to the previous finding, the blocked cue learning advantage was not replicated under 50% partial reinforcement across three experiments that 1) used the exact same training schedule as Uengoer et al. (2013), 2) additionally controlled the overall outcome rate at 50%, or 3) additionally conveyed base rate information of either 0% or 50% via cue absent trials. The redundancy effect was, however, evident when some uncertainty was introduced to the outcome of both redundant cues. These results prompted further exploration of the core mechanisms that underlie learning about redundant cues not only in standard conditions but also in probabilistic settings. Computational modelling of associative learning theories supports a primary role for the summed error learning algorithm in accounting for the overall results when some commonalities are assumed to exist among training and testing cues (Vogel & Wagner, 2017). The capability may be further improved by incorporating a preferential attention mechanism in favor of reliable predictors (Mackintosh 1975; Uengoer et al., 2020). Uengoer et al. (2020) originally proposed relative informativeness and absolute relationship as two antithetical theoretical contrasts derived from the different training schedules of the redundant cues. Alongside these two concepts, I argue that *outcome predictability* is another important formal quality of the learning context. These three concepts together provide an intuitive account of the conditions under which the redundancy

effect occurs, one which shares some similarities with formal associative accounts, which are differentially sensitive to these properties.

According to many theories, causal judgment involves effortful, controlled reasoning processes that are essential to human cognitive competence (Mitchell et al., 2009). This is obviously relevant to the way in which people reason about causally ambiguous cues. Motivated by the hypothesis that the redundancy effect may be driven by differences in reasoning about the blocked and uncorrelated cues, the experiments in Chapter 3 manipulated prior assumptions hypothesised to affect causal deduction. In doing so, these experiments found strong evidence that the judgment difference between the two cues depends on whether assumptions held permit valid deductions of non-causality. The blocked cue can be confidently inferred as non-causal if one entertains additive assumptions about outcome magnitude but remains causally ambiguous when non-additive assumptions are highlighted. Similarly, the non-causal nature of the uncorrelated cue can be deduced if one possesses non-preventative assumptions but not if preventative relationships are considered possible. In accord with these hypotheses, the redundancy effect was observed following pretraining that encouraged non-additive assumptions (i.e. elevating likelihood that the blocked cue is causal) and pretraining that encouraged non-preventative assumptions (i.e. lowering the likelihood that the uncorrelated cue is causal), but diminished among additive or preventative participants. These results suggest that the readiness to draw deductive inferences hinges on existing beliefs about how multiple putative causes combine. Difference in the ease with which deduction of non-causality is applied then drives judgment difference between the blocked and uncorrelated cues, giving rise to the redundancy effect when the propensity to engage in logical reasoning is stronger for the uncorrelated cue.

While these findings are consistent with a propositional view of human learning, the redundancy effect cannot be concluded to result solely from reasoning processes. It was

assumed that engagement with deductive inference would allow learners to confidently remove the possibility that the redundant cues are causal. This behavioural pattern is identifiable by low likelihood judgment combined with high confidence judgment. Clustering analyses of individual variability in the tendency to exhibit this dissociation showed that, even though a subset of participants engaged in confident elimination for the uncorrelated cue in conditions that elicited a significant redundancy effect, there was a considerable proportion of participants that did not. Likewise, a small proportion demonstrated confident elimination for the blocked cue in groups that abolished the redundancy effect, but the remaining individuals did not show signs of such processes. More importantly, the redundancy effect was found among participants who demonstrated no evidence of confident elimination at all for either cue, suggesting that alternative processes must also be at work to favor learning about the blocked cue over the uncorrelated cue. For example, basic memory encoding and retrieval mechanisms from an associative perspective (Thorwart & Livesey, 2016) may bias learning toward the blocked cue, which was more consistently paired with the outcome over the partially reinforced uncorrelated cue. While it is unclear how the operation of this fundamental form of learning may interact with higher order reasoning to influence judgement of ambiguous cues in humans, there is enough evidence to warrant considering a multiprocess theory for the redundancy effect.

Further convincing evidence for the involvement of propositional reasoning came from Chapter 5, where presenting the contingency information in the form of a written summary instead of the actual experience of training trials was sufficient to yield clear evidence of the redundancy effect. In fact, the relative validity effect was observed much more clearly in the summary-form Experiment 5.2 than in most of the contingency learning experiments that tested for this comparison in Chapters 2 and 3. Both effects were considerably enhanced under the food allergist scenario in which deductive inference about the uncorrelated cue was

encouraged compared to the hormone change scenario in which the use of deduction was discouraged. These findings highlight the influence of prior knowledge about cause and effect, suggesting that the sets of assumptions that participants formulate as a consequence of their everyday real-world interactions can have an impact on judgement about redundant cues beyond the explicit manipulations enforced by the laboratory environment (Greenaway & Livesey, 2020).

The involvement of inferential reasoning in the redundancy effect prompted a further question as to whether future learning about the blocked and uncorrelated cues would continue to be driven by higher order cognitive processes. Chapter 4 set out a series of experiments to test the adequacy of the theory protection principle (Spicer et al., 2020; 2022) within the context of the redundancy effect. The use of cognitive heuristic to protect existing relationships in a causal network is congenial to the reliance on deductive inferences to arrive at a causal judgment, in the sense that both entail rational thinking and reasoning. By implementing a similar assumption manipulation (pretraining) phase prior to the redundancy effect training, likelihood judgment of the uncorrelated cues being causal was substantially reduced following non-preventative pretraining compared to preventative pretraining.

The impact of magnitude additivity assumptions on the blocked cues was less compelling. When administered on their own, additive pretraining reduced likelihood judgement about the blocked cues compared to non-additive pretraining. However, this reduction in likelihood ratings for the blocked cues appeared to be accompanied by a simultaneous reduction in likelihood ratings for the uncorrelated cues, resulting in no observable change in the redundancy effect. In view of previous literature (Jones et al., 2019) that demonstrated a dissociation between likelihood judgment and causal certainty, one might expect the higher likelihood ratings to the blocked cues than to the uncorrelated cues following both additive pretraining and non-additive pretraining to correspond to lower

confidence for the blocked cues than for the uncorrelated cues. According to theory protection (Spicer et al., 2020), this dissociation might be further expected to lead to an enduring learning bias toward the blocked cue during compound training. These predictions were not borne out by the results from Experiment 4.3. Rather, the redundancy effect in both groups did not elicit an inversely correlated confidence pattern or guide subsequent bias in compound training.

These findings suggest that the mere presence of the redundancy effect prior to compound training is not sufficient to drive a subsequent learning bias toward the blocked cue. Instead, the magnitude of the difference in causal judgement after Stage 1 learning may be more important for future learning. Non-preventative pretraining in Experiments 4.1, 4.2, and 4.4 permitted deductive reasoning about the uncorrelated cues. Across these conditions, the redundancy effect was significantly enhanced by the inference that the uncorrelated cue must be non-causal. On this interpretation, it is the propensity to protect the deduced non-causal inference about the uncorrelated cue from violation, together with the concurrent presence of a blocked cue that was more consistent with the outcome, that leads to the biased attribution of the outcome to the blocked cue. On the whole, these findings suggest that existing assumptions and beliefs can have an enduring impact on how human learners derive meaning from ambiguous causes.

The following sections will delve into three issues of relevance to various aspects of the research and the general conclusions of the thesis as a whole. First, Chapter 2 identified models that combined a summed error term with a common element as the best performing models for explaining findings related to the redundancy effect across a number of probabilistic conditions. The modulating influence of the common cue is, however, not a constant factor but rather varies depending on both the overall outcome rate and the outcome probability of the target cue. Second, although Chapter 3 revealed judgment patterns that

convincingly suggest the involvement of deductive reasoning among a subset of participants, it is not unreasonable to assume that simpler associative mechanisms operate as a foundation for the biased learning toward the blocked cue, especially since the redundancy effect was still evident in conditions that were less amenable to deduction about the uncorrelated cue, such as those in Chapter 2. Third, Chapter 4 found that existing beliefs may have a long-lasting impact on the subsequent learning of redundant cues. These results are subject to alternative interpretations from error-driven attention theories.

Role of the Common Context

An intuitive explanation proposed in Pearce et al. (2012) and Uengoer et al. (2013) is that the blocked cue holds higher relative informativeness than the uncorrelated cue. This simple explanation affords particular theoretical value under conditions of partial reinforcement where outcome probability and predictability are at least partly controlled for between the critical cues. However, analysis of relative informativeness is usually somewhat simplistic in that it is made solely with reference to the discrete cues. In contrast, the error correction approach to modelling the redundancy effect benefits greatly from including a common element representing shared and/or contextual elements of the learning trial (Vogel & Wagner, 2017). In fact, we did not formally evaluate error correction models *without* this common element in Chapter 2 because we did not consider them justifiable. This is important because if learning of context is taken into account, then the predictive utility of the redundant cues would not only need to be compared against the cues accompanying them, but would also need to be compared against the common context cue present on every training and test trial.

Let us consider specifically a comparison of the critical cues (the blocked and uncorrelated cues) and the context. The base rate of the outcome across training establishes the informativeness of the context. Depending on the overall rate of outcome occurrence, the

context becomes established as a stronger, weaker, or equivalent outcome predictor that modulates likelihood judgment of the redundant cues. The base rates for the blocking and relative validity procedures establish the informativeness of the redundant cues. In a typical deterministic design, p(outcome|cue) is usually closer to the base rate for the uncorrelated cue than it is for the blocked cue. Indeed, outcome probability for the uncorrelated cue is exactly the same as that for the context if the overall outcome rate is controlled at 50%. The degree to which the blocked cue provides useful information above the context, and hence the degree to which its redundancy is reduced relative to the context, depends on its ability to signal an increase or decrease in outcome probability that is different to that predicted by the context. If also assuming 50% base rate in this instance, the 100% trained blocked cue signals a 50 percentage-point increase in outcome rate. Thus, under the standard condition, the informativeness of the blocked cue relative to the context is higher than that for the uncorrelated cue.

The situation is more complex when the blocking contingencies involve partial reinforcement. For partially reinforced cues, a better predictor on reinforced trials is simultaneously a poorer predictor on non-reinforced trials. With equated proportion of outcome-present and outcome-absent trials, the chance level accuracy cannot be improved or impaired by considering additional cues. As a consequence, both the blocked cue and the uncorrelated cue are just as informative as the context under 50% reinforcement rate. On the other hand, if the proportion of outcome-present and outcome-absent trials is imbalanced, then concurrent cues may vary in the absolute accuracy with which their outcomes can be predicted. For example, this is the case if the blocking treatment comprises 80% of outcome occurrences while the relative validity design comprises 80% of outcome occurrences on canonically reinforced trials and 0% on non-reinforced trials (i.e. the design of Experiment 2.4 in Chapter 2). The context, which predicts the outcome on 50% of the occasions based on

the overall outcome rate, acts as a poorer predictor than the blocked cue on 80% of trials involving the blocked cue but a better predictor on the remaining 20%. In other words, the context incorrectly signaled a 30 percentage-point drop in outcome probability following the blocked cue. Notably, however, this is not the case for relative validity where the uncorrelated cue predicts the outcome better than the context by 10% on non-reinforced trials but less well by 10% on reinforced trials. The predictive utility of the uncorrelated cue was balanced out across the two trial types and thus remains equivalent to that of the context.

The differential susceptibility to base rate and outcome probability variations suggests that the informativeness of the blocked and uncorrelated cues relative to the context may contribute to differential learning about the two cues. It could be argued that the reduction in the informational value of the blocked cue from being more informative than the context under 100% consistent reinforcement to being equally informative as the context under 50% intermittent reinforcement diminished the blocked cue learning advantage in Experiments 2.1 and 2.2, and the redundancy effect was restored when the higher relative informativeness of the blocked cue was regained in the 80% reinforcement condition of Experiment 2.4. These findings leave open the possibility that the redundancy effect results from the better informativeness of the blocked cue than the context. Future research is welcome to test this hypothesis.

Basing evaluation of relative informativeness on the comparison between redundant cues and the context is not without flaws. Even though the uncorrelated cue remained as equivalent predictors as the context in all four experiments in Chapter 2, its informativeness was unarguably lower than the more reliable predictors B and C that accompanied it. The blocked cue, which has never been trained together with more informative cues, should in any case be treated as a more likely cause than the uncorrelated cue. However, moving beyond the effects of relative informativeness on learning, it is debatable whether relative

informativeness also plays a part in retrieval-at-test effects. It could be the case that companion cues at test are weighed more heavily than those that were compounded during training but require additional retrieval effort at test. That is, comparison of each redundant cue with the context may contribute more significantly to likelihood judgment than comparison with cues that accompanied them during training but were absent at test.

Associative Memory as a Foundational Mechanism

Chapters 3, 4 and 5 suggest that human learners are capable of forming proposition-based inferences about causal relationships between ambiguous environmental cues. However, there are circumstances in which such complex mental ability is unlikely to be executed. For example, clustering analysis in Chapter 3 found the redundancy effect within a subset of participants who did not adopt confident elimination in their judgment process for both the blocked and uncorrelated cues. Assuming that the dissociative judgment pattern between causal likelihood and prediction certainty is indicative of deductive reasoning, it would be unlikely that these participants arrived at the learning bias using complex decision strategies rather than simpler learning and memory mechanisms. However, there might not always be a perfect mapping between the decision process and the introspection process such that individuals can still engage in deductive reasoning without showing the typical dissociation. It would also be contentious to claim that the redundancy effect observed in rats and pigeons (Pearce et al., 2012) is a result of similar deductive reasoning processes. Therefore, it is reasonable to assume that there are essential mechanisms that more fundamentally underlie the redundancy effect across a wide range of situations.

Associative memory, both its encoding and retrieval, is one such mechanism that works in concert with propositional reasoning to drive the redundancy effect especially when correct inferences are not readily allowed. On this view, differential learning follows directly from the fact the blocked cue is paired with the outcome with a higher probability than the

uncorrelated cue in the standard condition, and thus encodes the outcome more strongly during training and retrieves the outcome more strongly at test than the uncorrelated cue. Furthermore, Chapter 2 highlighted the capacity of a summed error model with common element to capture the pattern of learning observed across a range of redundancy effect designs. It is possible that associative memories of the common cue leading to the outcome also serves a primary role in guiding a learning bias toward the blocked cue. In the same way that the common cue overshadows the blocked and uncorrelated cues in terms of associative strength, associative memory pertaining to the co-presence of the common cue and the outcome may restrict the strength of memory independently attributed to the target cues. Rather than deductively infer the causal status of redundant cues from existing knowledge, exploration of cues that establish memories of outcome occurrence to various extents inductively yields outcome expectation differences.

Attentional Accounts as Alternatives to Theory Protection

Judgement of the uncorrelated cue as an unlikely cause reduced its utilisation in subsequent compound training in Experiments 4.1, 4.2 and 4.4. According to the theory protection principle, low likelihood judgement of the uncorrelated cue reflects the strong belief that the uncorrelated cue cannot be causal as a consequence of appropriate deduction. This belief is so strongly held that learners are resistant to updating the causal value of the uncorrelated cue when a different outcome is encountered in future. The concurrently presented blocked cue thus takes advantage of its ambiguous causal status and enters into a causal relationship with the new outcome. The selective attribution of the outcome to the blocked cue can be seen as being guided by the long-lasting influence of propositional reasoning which produces a tendency to protect existing theory about the uncorrelated cue.

The theory protection explanation predicates the redundancy effect on the typical dissociative judgment pattern between causal likelihood and prediction certainty. However,

as Chapter 4 has shown, lower likelihood judgement about the uncorrelated cue than for the blocked cue was not consistently accompanied by confidence judgement in the reverse direction. Although more sensitive measures might reveal subtle differences in confidence that were not detected in the current experiments, these findings are open to alternative interpretations from an associationist point of view that do not place particular emphasis on confidence.

Attentional theories predict lack of learning about the uncorrelated cue under the goal of error minimisation. Mackintosh's classic model of selective attention (1975) and Kruschke's EXIT model (2005) both express the idea that more processing resources are devoted to the cue that is better able to predict an outcome of relevance than less reliable predictors. During compound training, attending to the uncorrelated cue elicits the prediction of outcome absence, which generates a large discrepancy with the observed outcome presence. In an attempt to reduce prediction error, attention is shifted toward the moderately predictive blocked cue which, as a result, becomes more strongly associated with the outcome. As long as the redundant cues possess different associative strengths at the start of compound training, attention reallocation according to the predictive utility of each cue will yield asymmetrical learning within the compound.

The predictive power of the blocked and uncorrelated cues switches when further pairing of their compound involves outcome absence. In this case, more learning should take place about the uncorrelated cue as it is now a better predictor of the absence of the outcome. Note that this prediction is in keeping with that following from the theory protection principle, which expects strengthening of the existing knowledge that the uncorrelated cue is non-causal. However, the compound test procedure may be intrinsically flawed to gauge associative change during non-reinforced compound training in situations where preventative relationships are deemed implausible. For example, assuming that the outcome of a causal

cue is unpreventable may be required to establish the redundancy effect in the first place, but further training of the blocked-uncorrelated cue compound with outcome absence directly contradicts this assumption. This is where a potential caveat for the use of compound test procedure is warranted.

Limitations and Future Directions

The present thesis is broadly consistent with the view that learning about redundant cues entails the concurrent operation of complex inferential reasoning and simple association formation. While the applicability of deductive inferences has been shown to be highly dependent on the set of preconceived assumptions that learners entertain, and hence the causal context that presupposes these assumptions, it is not yet clear as to whether fundamental association-formation processes can interact with high-level cognition to also show sensitivity to assumptions implied in different causal scenarios. For instance, Uengoer and colleagues (2020) found the redundancy effect using a staged blocking design under the food allergist task. Based on the inconsistency between this finding and their simulation results, the authors rejected the adequacy of the common element approach to explain the redundancy effect. Likewise, the present set of studies were conducted without direct replication of Uengoer et al. (2013) using the food allergist task. In addition to the obvious need to replicate these findings in a task scenario that does not overly emphasise non-preventative relationships, it is important to also consider the degree of generalisation between cues under different circumstances. It may be the case that, under the dominant non-preventative assumption, cues that have led to outcome absence are seen as distinctive from consistently reinforced cues, and as such, the uncorrelated cue does not share much in common with the blocked cue. If this hypothesis were true, then Uengoer et al.'s (2020) finding can be successfully predicted by the common element model. It is also worth-noting that in most of the present experiments, instructions still noted that prevention is possible,

even though this point was not emphasised with pretraining or further instructions, as was the case in the preventative group. It is possible that the hormone change scenario did emphasise preventative cue interactions. Nevertheless, preventative pretraining clearly strengthened this further as the key comparison between the blocked and uncorrelated cues changed with the addition of preventative pretraining.

Base-rate manipulations in Chapter 2 yielded equivocal results regarding the exact contribution of the shared component to learning about ambiguous cues. Particularly, it remains to be solved how an epistemically uncertain blocked cue (i.e. consistently reinforced) and an aleatorily uncertain blocked cue (i.e. partially reinforced) are subject to the different influence of learning context. The idea that the cues do not necessarily share a unified common cue but a constellation of overlapping features or elements (McLaren & Mackintosh, 2000, 2002; Harris, 2006; Harris & Livesey, 2010) may allow further insights into the differential learning about redundant cues trained under different overall outcome rate and with different outcome predictability. Given that the strength of association between the overlapping components and the outcome relies critically on the underlying base rate, future research is warranted to conduct systematic manipulation of overall outcome rate in different causal contexts.

Attention deployment in favor of good predictors may guide learning about redundant cues, but based on computational modeling conducted in Chapter 2, its involvement appears to be secondary. This finding may partially explain previous failures to find differential alpha effects for the blocked and uncorrelated cues (Jones & Zaksiate, 2018; Uengoer et al., 2019). However, error-driven redistribution of attention may have a more important role to play in determining the fate the redundant cues in the long run. Monitoring attention allocation to the redundant cues during a compound test procedure would allow further insights into the varying influence of attention in ambiguous learning situations.

We note a companion issue with the simplified version of the hormone change scenario. Jones and Zaksaite (2020) allowed hormone levels to increase, decrease or stay the same in their original cover story. To match the conventional binary outcome in causal learning tasks, the current experiments removed 'hormone decrease' as an option on the likelihood judgment scale. However, the dimensional construct 'hormone level' implies in itself the possibility that changes can take place in both directions. And because the neutral point is not set at zero but at the centre of the scale, excitatory and inhibitory influence may be hard to distinguish on single cue trials. However, this should not change the interpretation regarding the preventative role of B where relative changes in the causal strength of Y were of most concern. Future research can benefit from using a bidirectional scale that includes either hormone decrease or prevention of a hormone increase to explore preventative relationships further. In addition, labelling the scale explicitly in terms of causal strength of the cue in question, rather than likelihood, might help differentiate causal beliefs from confidence.

Aside from potential improvements to the test structure, future studies would further explore ways to design more powerful and sensitive tests. The confidence ratings scale provides a straightforward way of estimating causal certainty. However, the phrasing of the instructions may not be as clear as it is intended. It remains a possibility that learners' self-reported confidence is not a true reflection of causal certainty but a derivative of likelihood judgment. The confusion may be resolved by giving illustrative examples clearly distinguishing self-reported confidence from perceived causal likelihood. Moreover, it is a commonplace to treat causal expectation elicited by a cue compound as an algebraic sum of its parts (Waldmann, 2007). This kind of linear combination rule may be undermined if cues are considered capable of preventing an outcome from occurring during non-reinforced compound training. Manipulations that undermine the additive combination rule may lead to

unexpected and uninterpretable results from the compound test procedure. Further examination of the malleability of such rules is therefore warranted.

Conclusion

The redundancy effect represents a theoretical conundrum in human causal learning literature. In an attempt to understand the processes by which human learners make causality judgement under ambiguity, the current thesis evaluated candidate explanations from the perspective of both elementary association formation and complex propositional reasoning. The overall findings suggest that an integrated explanation is needed to account for the various circumstances that influence causal judgment of redundant cues. These factors include existing knowledge, outcome uncertainty, task scenario, as well as individual variability in the tendency to engage in different learning processes. While simulation results suggest that the summed error learning algorithm (e.g. as used in the Rescorla-Wagner model) is the primary driver for the redundancy effect when an overlapping component is assumed to be shared among cues, effortful and controlled cognition that enables logical inferences may take precedence under appropriate assumptions and exert a far-reaching impact on future learning about redundant cues.

References

Baetu, I., Baker, A. G., Darredeau, C., & Murphy, R. A. (2005). A comparative approach to cue competition with one and two strong predictors. *Learning & Behavior*, *33*, 160–171. https://doi.org/10.3758/BF03196060

Barberia, I., Blanco, F. & Rodríguez-Ferreiro, J. (2021). The more, the merrier: Treatment frequency influences effectiveness perception and further treatment choice. *Psychon Bull Rev, 28*, 665–675. https://doi.org/10.3758/s13423-020-01832-6

Barbey, A., & Barsalou, L. (2009). Reasoning and Problem Solving: Models. In *Encyclopedia of Neuroscience* (Vol. 8, pp. 35–43).

Bates, L. A., Sayialel, K. N., Njiraini, N. W., Poole, J. H., Moss, C. J., & Byrne, R. W. (2008). African elephants have expectations about the locations of out-of-sight family members. *Biology letters*, *4*(1), 34–36. https://doi.org/10.1098/rsbl.2007.0529

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

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of experimental psychology. General*, *135*(1), 92–102. https://doi.org/10.1037/0096-3445.135.1.92

Boddez, Y., Vervliet, B., Baeyens, F., Lauwers, S., Hermans, D., & Beckers, T. (2012). Expectancy bias in a selective conditioning procedure: Trait anxiety increases the threat value of a blocked stimulus. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(2), 832–837. https://doi.org/10.1016/j.jbtep.2011.11.005

Bradfield, L., & McNally, G. P. (2008). Unblocking in Pavlovian fear conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 34*(2), 256–265. https://doi.org/10.1037/0097-7403.34.2.256

Callejas-Aguilera, J. E., & Rosas, J. M. (2010). Ambiguity and context processing in human predictive learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(4), 482–494. https://doi.org/10.1037/a0018527

Chan, Y. Y., Westbrook, R. F., & Holmes, N. M. (2021). Protecting the Rescorla–Wagner (1972) theory: A reply to Spicer et al. (2020). *Journal of Experimental Psychology: Animal Learning and Cognition, 47*(2), 211–215. https://doi.org/10.1037/xan0000271

Chow, J. Y. L., Lee, J. C., & Lovibond, P. F. (2024). Using unobserved causes to explain unexpected outcomes: The effect of existing causal knowledge on protection from extinction by a hidden cause. *Journal of experimental psychology. Learning, memory, and cognition*, *50*(7), 1167–1185. https://doi.org/10.1037/xlm0001306

Cole, R. P., Barnet, R. C., Miller, R. R., & Hulse, S. H. (1995). Effect of Relative Stimulus Validity: Learning or Performance Deficit? *Journal of Experimental Psychology: Animal Behavior Processes*, *21*(4), 293–303. https://doi.org/10.1037/0097-7403.21.4.293

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*, *55*(3), 965–985. https://doi.org/10.1080/02724980143000578

De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the role of higher order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior*, *33*(2), 239–249. https://doi.org/10.3758/BF03196066

Denniston, J. C., Savastano, H. I., Miller, R. R., Mowrer, R. R., & Klein, S. B. (2001). The Extended Comparator Hypothesis: Learning by Contiguity, Responding by Relative Strength. In *Handbook of Contemporary Learning Theories* (1st ed., pp. 65–117). Psychology Press. https://doi.org/10.4324/9781410600691-3

Dickinson, A. (2001). The 28th Bartlett Memorial Lecture Causal learning: An associative analysis. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *54*(1), 3–25. https://doi.org/10.1080/02724990042000010

Dickinson, A., & Burke, J. (1996). Within compound Associations Mediate the Retrospective Revaluation of Causality Judgements. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *49*(1), 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 Section A*, *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., Goldwater, M. B., Greenaway, J. K., Hutchings, R., & Livesey, E. J. (2020). Relational rule discovery in complex discrimination learning. *Journal of experimental psychology. Learning, memory, and cognition*, *46*(10), 1807–1827. https://doi.org/10.1037/xlm0000848

Don, H. J., & Livesey, E. J. (2017). Effects of outcome and trial frequency on the inverse base-rate effect. *Memory & Cognition, 45*(3), 493–507. https://doi.org/10.3758/s13421-016-0667-y

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*(4), 1142–1163. https://doi.org/10.3758/s13423-020-01870-0

Durlach, P. J., & Rescorla, R. A. (1980). Potentiation rather than overshadowing in flavor-aversion learning: An analysis in terms of within-compound associations. *Journal of Experimental Psychology. Animal Behavior Processes*, *6*(2), 175–187. https://doi.org/10.1037/0097-7403.6.2.175

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society. B, Biological Sciences*, *278*(1718), 2553–2561. https://doi.org/10.1098/rspb.2011.0836

Gluck, M. A., & Bower, G. H. (1988). From Conditioning to Category Learning: An Adaptive Network Model. *Journal of Experimental Psychology. General*, *117*(3), 227–247. https://doi.org/10.1037/0096-3445.117.3.227

Greenaway, J. K., & Livesey, E. J. (2020). Can We Set Aside Previous Experience in a Familiar Causal Scenario? *Frontiers in psychology*, *11*, 578775. https://doi.org/10.3389/fpsyg.2020.578775

Hanus, D., & Call, J. (2011). Chimpanzee problem-solving: contrasting the use of causal and arbitrary cues. *Animal Cognition*, *14*(6), 871–878. https://doi.org/10.1007/s10071-011-0421-6

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological Review, 113*(3), 584–605. https://doi.org/10.1037/0033-295X.113.3.584

Harris, J. A., & Livesey, E. J. (2010). An attention-modulated associative network. *Learning & Behavior, 38*(1), 1–26. https://doi.org/10.3758/LB.38.1.1

Haselgrove, M. (2010). Reasoning Rats or Associative Animals? A Common-Element Analysis of the Effects of Additive and Subadditive Pretraining on Blocking. *Journal of Experimental Psychology. Animal Behavior Processes*, *36*(2), 296–306. https://doi.org/10.1037/a0016603

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1191–1206. https://doi.org/10.1037/a0013025

Holmes, N. M., Chan, Y. Y., & Westbrook, R. F. (2019). A Combination of Common and Individual Error Terms Is Not Needed to Explain Associative Changes When Cues With Different Training Histories Are Conditioned in Compound: A Review of Rescorla's Compound Test Procedure. *Journal of Experimental Psychology. Animal Learning and Cognition*, *45*(2), 242–256. https://doi.org/10.1037/xan0000204

Holland, J., Holyoak, K., Nisbett, R., & Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press. http://doi.org/10.1109/MEX.1987.4307100

Johansen, M. K., Fouquet, N., & Shanks, D. R. (2007). Paradoxical effects of base rates and representation in category learning. *Memory & Cognition*, *35*(6), 1365–1379. https://doi.org/10.3758/BF03193608

Jones, P. M., & Pearce, J. M. (2015). The fate of redundant cues: Further analysis of the redundancy effect. *Learning & Behavior*, *43*(1), 72–82. https://doi.org/10.3758/s13420-014-0162-x

Jones, P. M., & Zaksaite, T. (2018). The redundancy effect in human causal learning: No evidence for changes in selective attention. *The Quarterly Journal of Experimental Psychology, 71*, 1748–1760. https://doi.org/10.1080/17470218.2017.1350868

Jones, P. M., Zaksaite, T., & Mitchell, C. J. (2019). Uncertainty and blocking in human causal learning. *Journal of Experimental Psychology: Animal Learning and Cognition, 45,* 111–124. https://doi.org/10.1037/xan0000185

Kahneman, D., & Tversky, A. (1982). Intuitive prediction: Biases and corrective procedures. In *Judgment under Uncertainty* (pp. 414–421). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.031

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

Karazinov, D. M., & Boakes, R. A. (2007). Second-order conditioning in human predictive judgements when there is little time to think. *The Quarterly Journal of Experimental Psychology, 60*(3), 448–460. https://doi.org/10.1080/17470210601002488

Kattner, F. (2015). Transfer of absolute and relative predictiveness in human contingency learning. *Learning & Behavior, 43*(1), 32–43. https://doi.org/10.3758/s13420-014-0159-5

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1-16. http://dx.doi.org/10.1068/v070821

Kozyreva, A., & Hertwig, R. (2021). The interpretation of uncertainty in ecological rationality. *Synthese*, *198*(2), 1517-1547. https://doi.org/10.1007/s11229-019-02140-w

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

Lee, J. C., & Livesey, E. J. (2012). Second-order conditioning and conditioned inhibition: influences of speed versus accuracy on human causal learning. *PloS one*, *7*(11), e49899. https://doi.org/10.1371/journal.pone.0049899

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. B, Comparative and Physiological Psychology*, *57*(3), 193–243. https://doi.org/10.1080/02724990344000141

Le Pelley, M. E., Turnbull, M. N., Reimers, S. J., & Knipe, R. L. (2010). Learned predictiveness effects following single-cue training in humans. *Learning & Behavior*, *38*(2), 126–144. https://doi.org/10.3758/LB.38.2.126

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 56B*(1), 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

Leung, H. T., & Westbrook, R. F. (2010). Increased spontaneous recovery with increases in conditioned stimulus alone exposures. *Journal of Experimental Psychology: Animal Behavior Processes, 36*(3), 354–367. https://doi.org/10.1037/a0017882

Livesey, E. J., & Boakes, R. A. (2004). Outcome additivity, elemental processing and blocking in human causality judgements. *The Quarterly Journal of Experimental Psychology*, *57*(4b), 361–379. https://doi.org/10.1080/02724990444000005

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

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., Thorwart, A., De Fina, N. L., & Harris, J. A. (2011). Comparing learned predictiveness effects within and across compound discriminations. *Journal of experimental psychology. Animal behavior processes*, *37*(4), 446–465. https://doi.org/10.1037/a0023391

Lovibond, P. (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

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 (2006), *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), 276–298. https://doi.org/10.1037/h0076778

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. *Animal Learning & Behavior, 4*(2), 186–192. https://doi.org/10.3758/BF03214033

McLaren, I.P.L., Mackintosh, N.J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, *28*(3), 211–246. https://doi.org/10.3758/BF03200258

McLaren, I. P. L, & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior, 30*(3), 177–200. https://doi.org/10.3758/BF03192828

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *The Behavioral and Brain Sciences*, *32*(2), 183–198. https://doi.org/10.1017/S0140525X09000855

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 B: Comparative and Physiological Psychology, 55B*(4), 311–329. https://doi.org/10.1080/02724990244000025

Mitchell, C. J., Lovibond, P. F., & Condoleon, M. (2005). Evidence for deductive reasoning in blocking of causal judgments. *Learning and Motivation*, *36*(1), 77–87. https://doi.org/10.1016/j.lmot.2004.09.001

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. *Quarterly journal of experimental psychology (2006)*, *59*(5), 830–844. https://doi.org/10.1080/17470210500242847

Moran, P. M., Owen, L., Crookes, A. E., Al-Uzri, M. M., & Reveley, M. A. (2008). Abnormal prediction error is associated with negative and depressive symptoms in schizophrenia. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *32*(1), 116–123. https://doi.org/10.1016/j.pnpbp.2007.07.021

Morris, R., Griffiths, O., Le Pelley, M. E., & Weickert, T. W. (2013). Attention to Irrelevant Cues Is Related to Positive Symptoms in Schizophrenia. *Schizophrenia Bulletin*, *39*(3), 575–582. https://doi.org/10.1093/schbul/sbr192

Murphy, R. A., Baker, A. G., & Fouquet, N. (2001a). Relative validity effects with either one or two more valid cues in Pavlovian and instrumental conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 27*(1), 59–67. https://doi.org/10.1037//0097-7403.27.1.59

Murphy, R. A., Baker, A. G., & Fouquet, N. (2001b). Relative Validity of Contextual and Discrete Cues. *Journal of Experimental Psychology: Animal Behavior Processes, 27*(2), 137–152. https://doi.org/10.1037/0097-7403.27.2.137

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. *Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

O'Connell, S., & Dunbar, R. I. M. (2005). The perception of causality in chimpanzees (Pan spp.). *Animal Cognition*, *8*(1), 60–66. https://doi.org/10.1007/s10071-004-0231-1

Öhman, A., & Dimberg, U. (1978). Facial expressions as conditioned stimuli for electrodermal responses: A case of "preparedness"? *Journal of Personality and Social Psychology*, *36*(11), 1251–1258. https://doi.org/10.1037/0022-3514.36.11.1251

Olsson, A., & Phelps, E. A. (2004). Learned Fear of "Unseen" Faces after Pavlovian, Observational, and Instructed Fear. *Psychological Science*, *15*(12), 822–828. https://doi.org/10.1111/j.0956-7976.2004.00762.x

Pearce, J. M. (1987). A Model for Stimulus Generalization in Pavlovian Conditioning. *Psychological Review*, *94*(1), 61–73. https://doi.org/10.1037/0033-295X.94.1.61

Pearce, J. M., Dopson, J. C., Haselgrove, M., & Esber, G. O. R. (2012). The fate of redundant cues during blocking and a relative validity. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 167–179. https://doi.org/10.1037/a0027662

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

Perales, J. C., & Shanks, D. R. (2008). Driven by power? Probe question and presentation format effects on causal judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1482–1494. https://doi.org/10.1037/a0013509

Rescorla, R. A. (1969). Pavlovian conditioned inhibition. *Psychological Bulletin, 72*(2), 77–94. https://doi.org/10.1037/h0027760

Rescorla, R. A. (2000). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behavior Processes, 26*(4), 428–438. https://doi.org/10.1037/0097-7403.26.4.428

Rescorla, R. A. (2001). Unequal Associative Changes when Excitors and Neutral Stimuli are Conditioned in Compound. *The Quarterly Journal of Experimental Psychology Section B*, *54*(1b), 53–68. https://doi.org/10.1080/713932743

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. Projasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton Century Crofts.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking.* The MIT Press.

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., & Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, *8*(9), 1234–1240. https://doi.org/10.1038/nn1527

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24*(4), 405–415. https://doi.org/10.1037/0097-7403.24.4.405

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

Syakur, M. A., Khotimah, B.K., Rochman, E.M. S., & Satoto, B.D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, *336*, 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Tannenbaum, D., Fox, C. R., & Ülkümen, G. (2017). Judgment Extremity and Accuracy Under Epistemic vs. Aleatory Uncertainty. *Management Science*, *63*(2), 497–518. https://doi.org/10.1287/mnsc.2015.2344

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

Uengoer, M., Lotz, A., & Pearce, J. M. (2013). The fate of redundant cues in human predictive learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *39*, 323–333. https://doi.org/10.1037/a0034073

Uengoer, M., Dwyer, D., Koenig, S., & Pearce, J. (2019). A test for a difference in the associability of blocked and uninformative cues in human predictive learning. *Quarterly Journal of Experimental Psychology, 72*(2), 222–237. https://doi.org/10.1080/17470218.2017.1345957

Uengoer, M., Lachnit, H., & Pearce, J. M. (2020). The Role of Common Elements in the Redundancy Effect. *Journal of Experimental Psychology: Animal Learning and Cognition, 46*(3), 286–296. https://doi.org/10.1037/xan0000236

van den Akker, K., Schyns, G., & Jansen, A. (2018). Learned Overeating: Applying Principles of Pavlovian Conditioning to Explain and Treat Overeating. *Current Addiction Reports*, *5*(2), 223–231. https://doi.org/10.1007/s40429-018-0207-x

Vogel, E. H., & Wagner, A. R. (2017). A Theoretical Note in Interpretation of the "Redundancy Effect" in Associative Learning. *Journal of Experimental Psychology. Animal Learning and Cognition*, *43*(1), 119–125. https://doi.org/10.1037/xan0000123

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31*(2), 233–256. https://doi.org/10.1080/15326900701221231

Wagner, A. R., Logan, F. A., & Haberlandt, K. (1968). Stimulus Selection in Animal Discrimination Learning. *Journal of Experimental Psychology, 76*(2p1), 171–180. https://doi.org/10.1037/h0025414

Wasserman, E. A., (1990). Attribution of Causality to Common and Distinctive Elements of Compound Stimuli. *Psychological Science, 1*(5), 298–302.

Wu, M., & Cheng, P. W. (1999). Why Causation Need not Follow From Statistical Association: Boundary Conditions for the Evaluation of Generative and Preventive Causal Powers. *Psychological Science*, *10*(2), 92–97. https://doi.org/10.1111/1467-9280.00114

Zaksaite, T., & Jones, P.M. (2017). The redundancy effect in human causal learning: Evidence against a comparator theory explanation. *Proceedings of the 38th Annual meeting of the Cognitive Science Society* (pp. 3640–3645). Austin, TX: Cognitive Science Society.

Zaksaite, T., & Jones, P. (2020). The redundancy effect is related to a lack of conditioned inhibition: Evidence from a task in which excitation and inhibition are symmetrical. *Quarterly Journal of Experimental Psychology, 73*(2), 260–278. https://doi.org/10.1177/1747021819878430

Supplementary Materials: Chapter 2

Experiment Instructions

Training

In this experiment you are asked to imagine that you are a medical researcher who is interested in studying the effects of different medicines on hormone levels. Your task is to figure out whether the consumption of different medicines will result in no change or an increase in hormone levels.

Sometimes one medicine will be consumed and sometimes two medicines will be consumed together. Note that these medicine names are novel and complex but to make your task easier, each one starts with a unique letter.

On the following screens you will see the medicines that Patient X consumes and will be asked to predict whether hormone level will not change or increase by clicking the corresponding button. Then, you will be informed of the resulting hormone level change, if any.

At the beginning you will have to guess but by using the feedback provided your guesses should become more accurate. Accuracy is more important than speed for your answers; you may take as long as you like on each trial. Please attend to the severity of the hormone level change, as this would help you determine which medicines are causing the change.

Please note that only the presented information can help you. Your own personal knowledge or experience with medicines of similar names will NOT help you in this task. Consequently, try to use only the knowledge you have learned from the experiment to make your decisions.

Now you will see some new medicines that Patient X consumes and will be asked to predict whether hormone level will not change or increase by clicking the corresponding button. Then, you will be informed of the resulting hormone level change, if any.

Ratings Test

Next, you are asked to provide your final report. Medicines will be presented on the screen and your task is to use all of the information you have collected about Patient X to judge the likelihood that specific medicines or combinations of medicines will cause an increase in hormone levels. For each medicine, you will be asked to provide two judgments.

First, please rate the likelihood that the medicine (or combination of medicines) causes a hormone increase on a scale from "Definitely DOES NOT cause increase" to "Definitely DOES cause increase" by clicking anywhere on the rating scale provided.

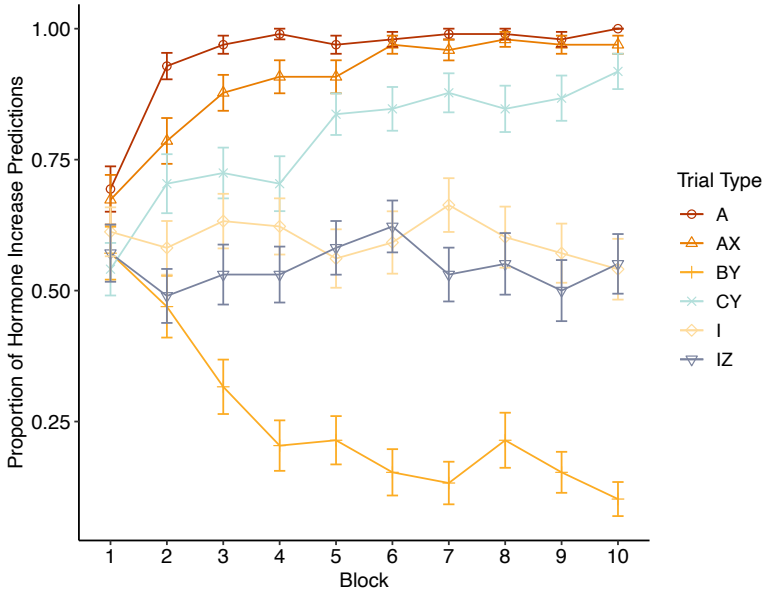
Second, please rate your confidence in your ability to judgment the likelihood that the medicine (or combination of medicines) causes a hormone increase, on a scale from "Not At All Confident" to "Very Confident" by clicking anywhere on the rating scale provided. Note that this judgment may be independent of the actual likelihood that you provided (regardless of whether you gave it a low, middle, or high likelihood rating, you may have reason to be more confident or less confident in your judgment).

Training Curves

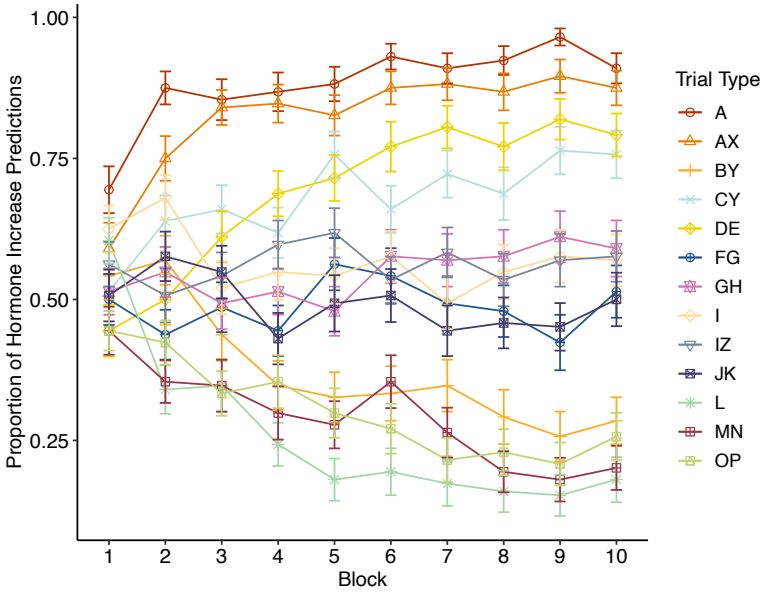
Figure S.1

Mean proportion of hormone increase predictions across ten blocks of training in (a) Experiment 2.1, (b) Experiment 2.2, (c) the 0% group of Experiment 2.3, (d) the 50% group of Experiment 2.3, and (e) Experiment 2.4. Error bars indicate standard error of mean (SEM).

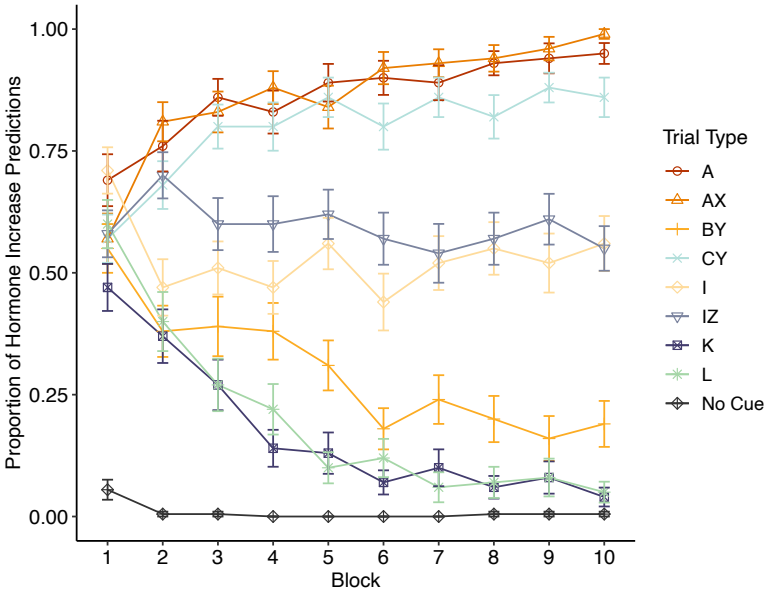
(a)



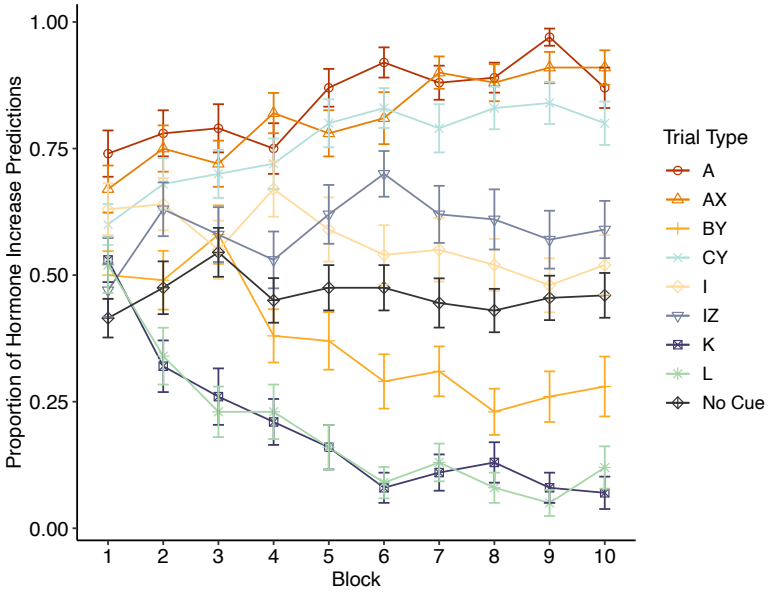
(b)



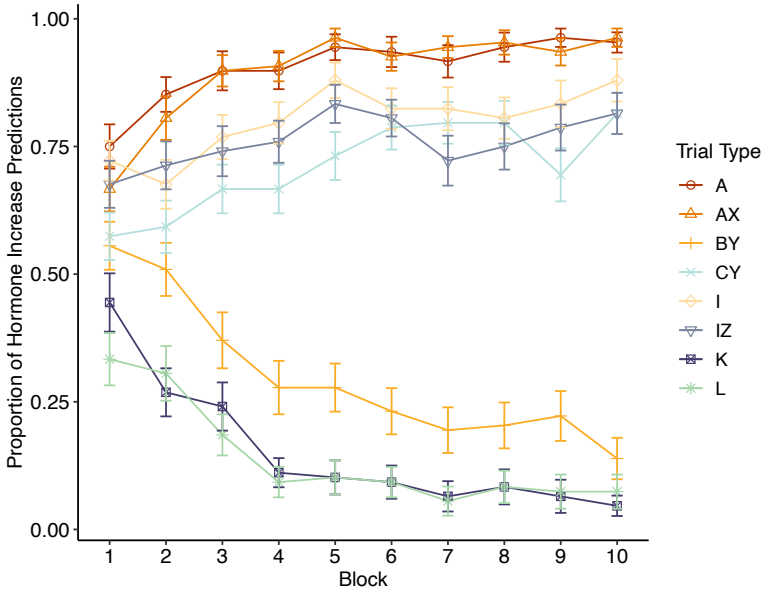
(c)



(d)



(e)



Ratings Test

Figure S.2

Mean likelihood ratings for all cues on the ratings test of Experiment 2.2. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

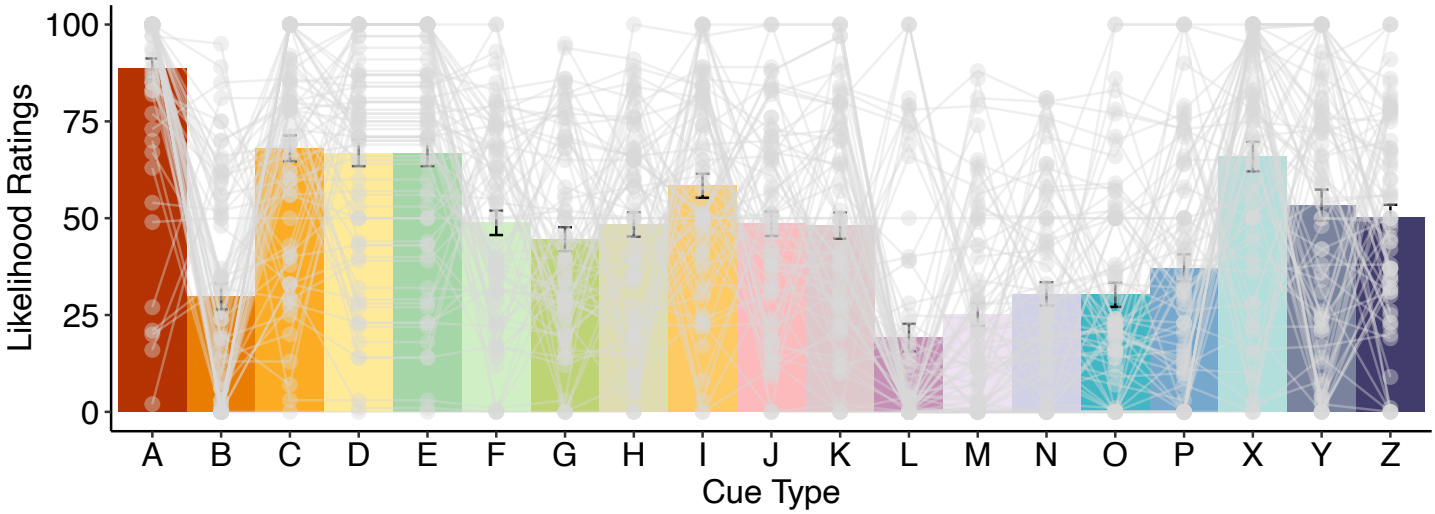


Figure S.3

Mean likelihood ratings for all cues on the ratings test of Experiment 2.3. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

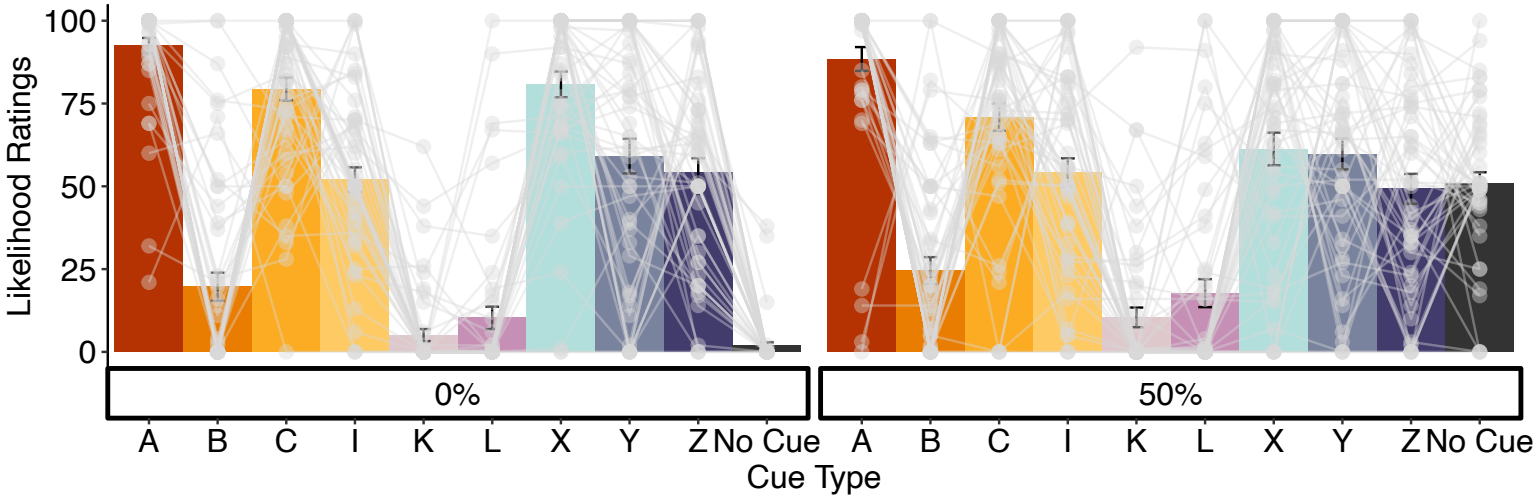
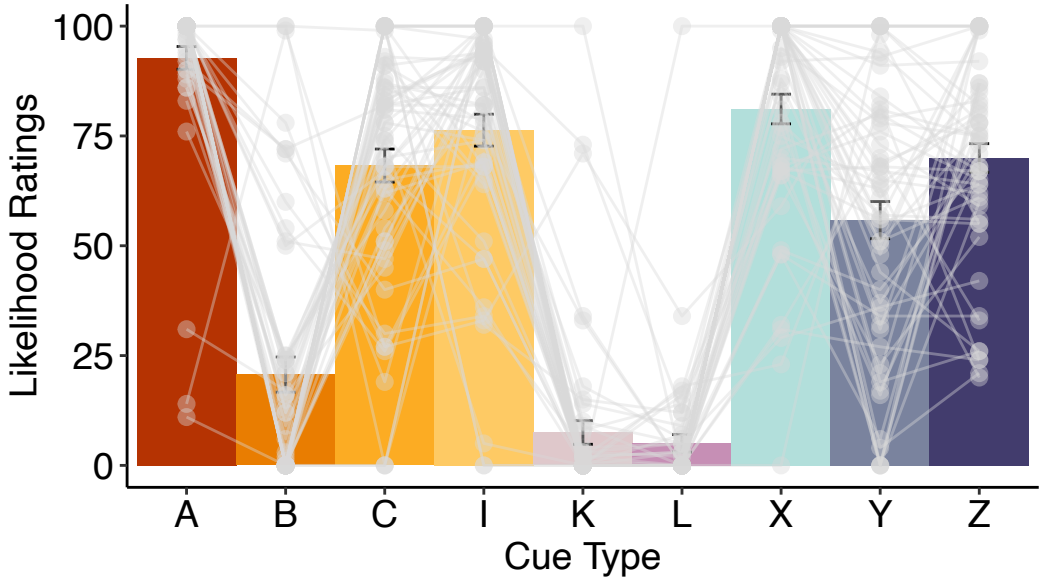


Figure S.4

Mean likelihood ratings for all cues on the ratings test of Experiment 2.4. Error bars indicate standard error of mean (SEM). Connected points represent data from the same particip



Model Fitting

Table S.1

Model fitting for Experiments 1-4 with training data only

Experiment	Model		Parameters						Fit statistics		
	Attention	Error term	Alpha	Beta	Theta	c1	c2	c3	- Log L	AIC	BIC
1	None	separable	0.728	0.222	-	0.666	1.000	4.756	2935.191	5880.381	5956.121
	None	summed	0.115	0.823	-	1.276	0.454	3.096	2436.208	4882.417	4958.156
	Mackintosh	separable	0.793	0.245	0.208	0.723	1.000	0.654	2895.918	5803.835	5894.723
	Mackintosh	summed	1.000	0.095	<.001	1.275	0.454	1.031	2436.198	4884.395	4975.283
	Uengoer	separable	0.657	0.509	0.389	0.758	0.997	1.252	2803.434	5618.867	5709.755
	Uengoer	summed	0.998	0.101	0.008	1.274	0.454	2.329	2433.895	4879.789	4970.677
2	None	separable	0.479	0.349	-	0.533	1.000	0.833	11288.034	22586.068	22673.39
	None	summed	0.587	0.151	-	0.953	0.437	4.896	10807.061	21624.121	21711.441
	Mackintosh	separable	0.412	0.537	0.020	0.511	1.000	3.986	11251.592	22515.184	22619.968
	Mackintosh	summed	0.984	0.105	0.123	0.957	0.432	4.590	10771.869	21555.739	21660.522
	Uengoer	separable	0.918	0.212	0.123	0.555	0.999	4.062	11256.774	22525.548	22630.332
	Uengoer	summed	0.884	0.115	0.014	0.949	0.436	3.429	10778.570	21569.139	21673.923
3	None	separable	0.247	0.692	-	0.759	1.000	2.355	10528.705	21067.411	21155.392
	None	summed	0.179	0.544	-	1.210	0.465	3.706	9780.102	19570.203	19658.185
	Mackintosh	separable	0.938	0.198	0.328	0.839	0.937	2.506	10301.633	20615.266	20720.844
	Mackintosh	summed	1.000	0.106	0.025	1.197	0.466	1.179	9758.566	19529.132	19634.710
	Uengoer	separable	0.986	0.257	0.301	0.812	0.995	3.439	10309.493	20630.987	20736.564
	Uengoer	summed	0.999	0.113	0.009	1.196	0.467	1.940	9744.951	19501.901	19607.479
4	None	separable	0.423	0.352	-	0.834	1.000	3.150	4013.052	8036.105	8115.506
	None	summed	0.291	0.398	-	1.245	0.454	4.148	3456.814	6923.628	7003.029
	Mackintosh	separable	1.000	0.173	0.399	0.959	0.967	3.329	3860.047	7732.094	7827.375
	Mackintosh	summed	1.000	0.141	0.085	1.221	0.453	4.634	3425.161	6862.323	6957.604
	Uengoer	separable	0.995	0.269	0.463	0.941	0.967	4.884	3807.790	7627.580	7722.861
	Uengoer	summed	1.000	0.148	0.019	1.222	0.453	0.225	3425.114	6862.227	6957.509

Table S.2

Model fitting for Experiments 1-4 with cue choice data only

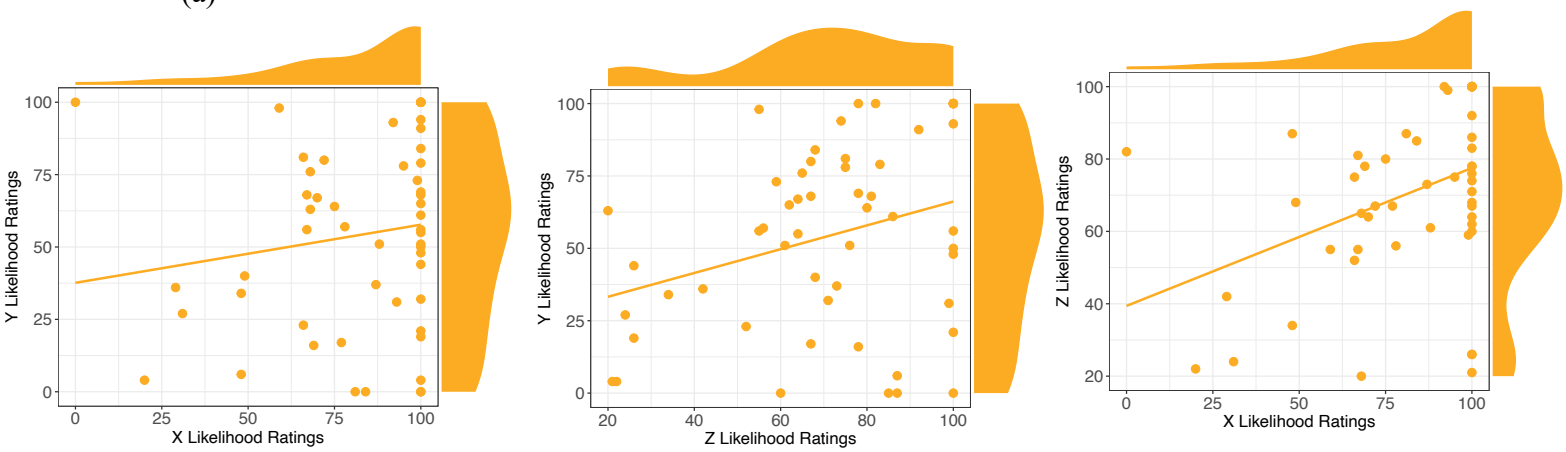
Experiment	Model		Parameters						Fit statistics		
	Attention	Error term	Alpha	Beta	Theta	c1	c2	c3	- Log L	AIC	BIC
1	None	separable	0.095	0.322	-	2.636	0.782	1.752	350.846	711.693	765.460
	None	summed	0.984	0.009	-	1.714	0.759	2.780	351.296	712.593	766.360
	Mackintosh	separable	0.965	0.065	0.120	3.073	1.000	2.131	348.755	709.510	774.031
	Mackintosh	summed	0.989	0.014	0.111	2.201	0.588	2.707	351.163	714.325	778.846
	Uengoer	separable	0.141	0.030	0.015	3.140	0.957	4.026	343.820	699.639	764.160
	Uengoer	summed	0.017	0.072	0.014	3.735	0.877	4.645	344.370	700.740	765.261
2	None	separable	0.032	0.918	-	0.877	0.578	1.458	1158.058	2326.116	2390.663
	None	summed	0.085	0.045	-	0.426	0.572	2.954	1160.871	2331.743	2396.290
	Mackintosh	separable	0.653	0.003	<.001	0.483	0.967	3.417	1164.906	2341.812	2419.269
	Mackintosh	summed	0.684	0.006	<.001	4.671	0.921	2.954	1160.871	2333.743	2411.200
	Uengoer	separable	0.731	0.040	<.001	0.600	0.401	1.459	1158.058	2328.116	2405.573
	Uengoer	summed	0.646	0.006	<.001	3.623	0.717	2.954	1160.871	2333.743	2411.199
3	None	separable	0.110	0.375	-	0.490	0.129	1.354	764.997	1539.994	1600.895
	None	summed	0.093	0.133	-	4.208	0.992	2.250	768.837	1547.675	1608.576
	Mackintosh	separable	1.000	0.074	0.138	3.691	0.501	2.369	750.917	1513.833	1586.914
	Mackintosh	summed	1.000	0.092	0.174	2.470	0.199	1.917	765.887	1543.774	1616.855
	Uengoer	separable	0.100	0.099	0.012	0.738	0.165	3.017	755.211	1522.422	1595.503
	Uengoer	summed	0.127	0.006	0.053	3.171	0.547	5.000	760.125	1532.250	1605.331
4	None	separable	0.017	0.646	-	3.115	0.047	2.987	368.793	747.585	802.137
	None	summed	0.015	0.136	-	3.100	0.808	4.463	369.105	748.211	802.763
	Mackintosh	separable	0.633	0.025	0.024	4.723	0.183	3.044	368.510	749.021	814.483
	Mackintosh	summed	0.985	0.004	0.295	2.793	0.840	5.000	364.700	741.393	806.856
	Uengoer	separable	0.368	0.023	0.003	2.478	0.228	3.268	368.693	749.385	814.848
	Uengoer	summed	0.649	0.003	<.001	2.943	0.763	4.462	369.105	750.211	815.673

Correlational Analysis for the Common Element Model

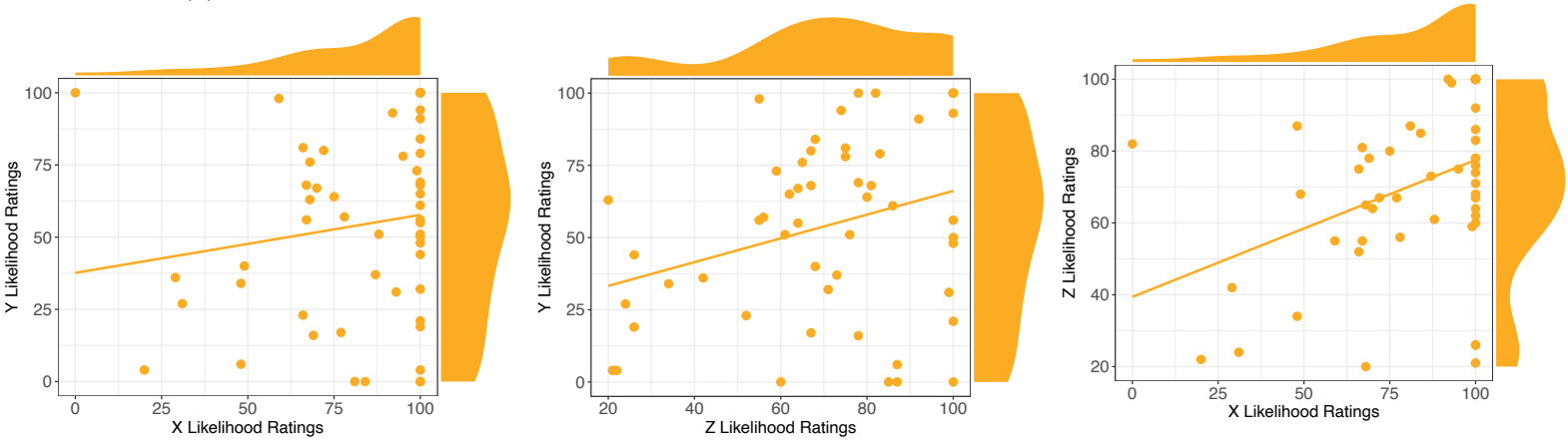
Figure S.5

X vs. Y, Z vs. Y, and X vs. Z correlations simulated under the summed error learning algorithm without attention across (a) Experiment 2.1, (b) Experiment 2.2, (c) Experiment 2.3, and (d) Experiment 2.4.

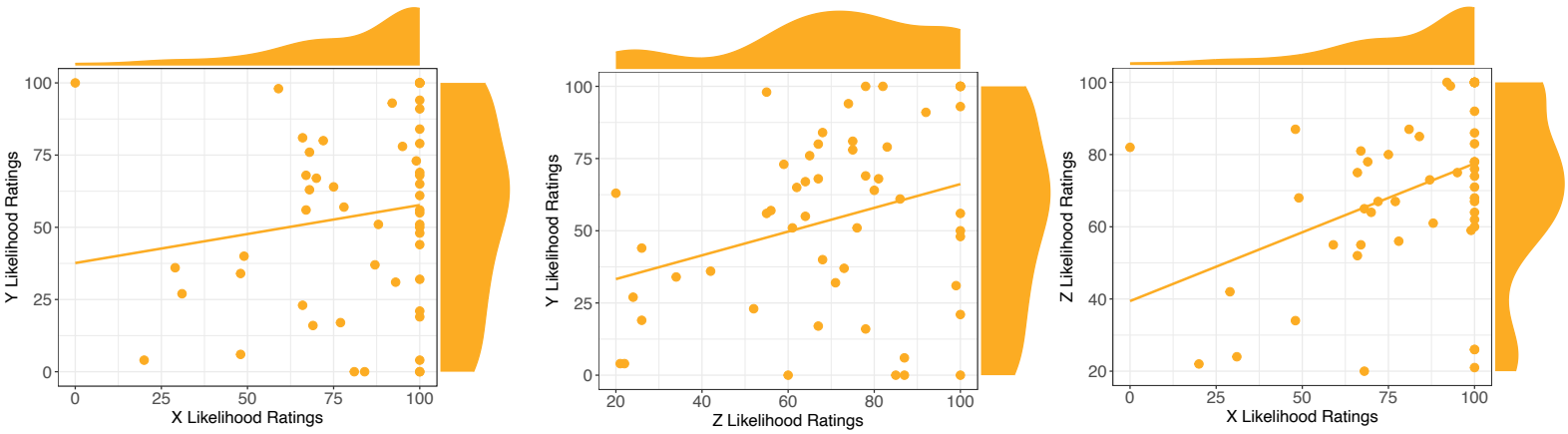
(a)



(b)



(c)



(d)

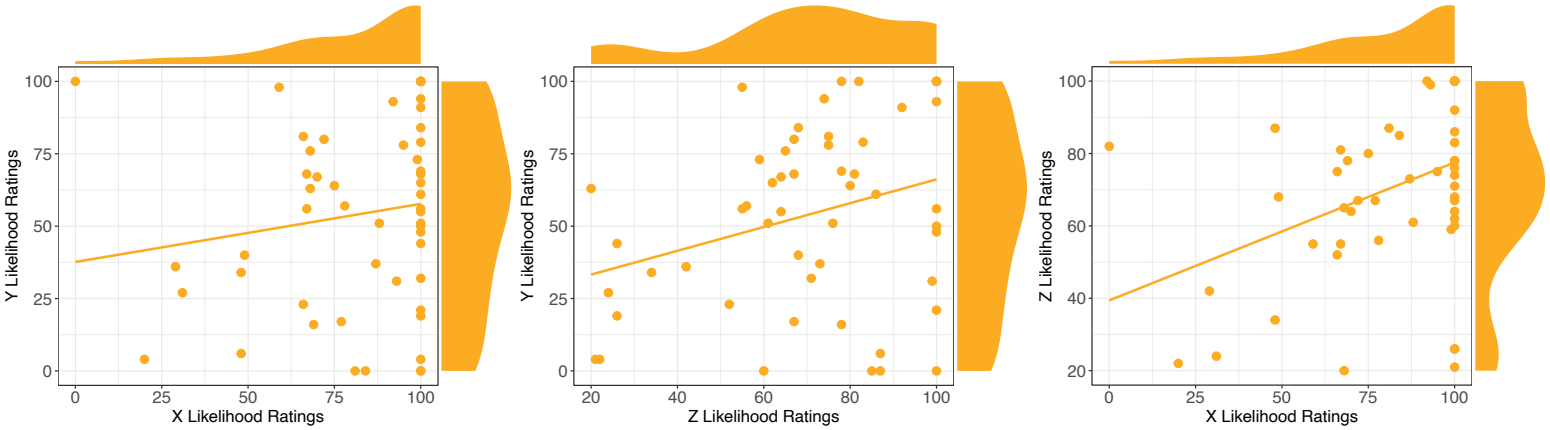
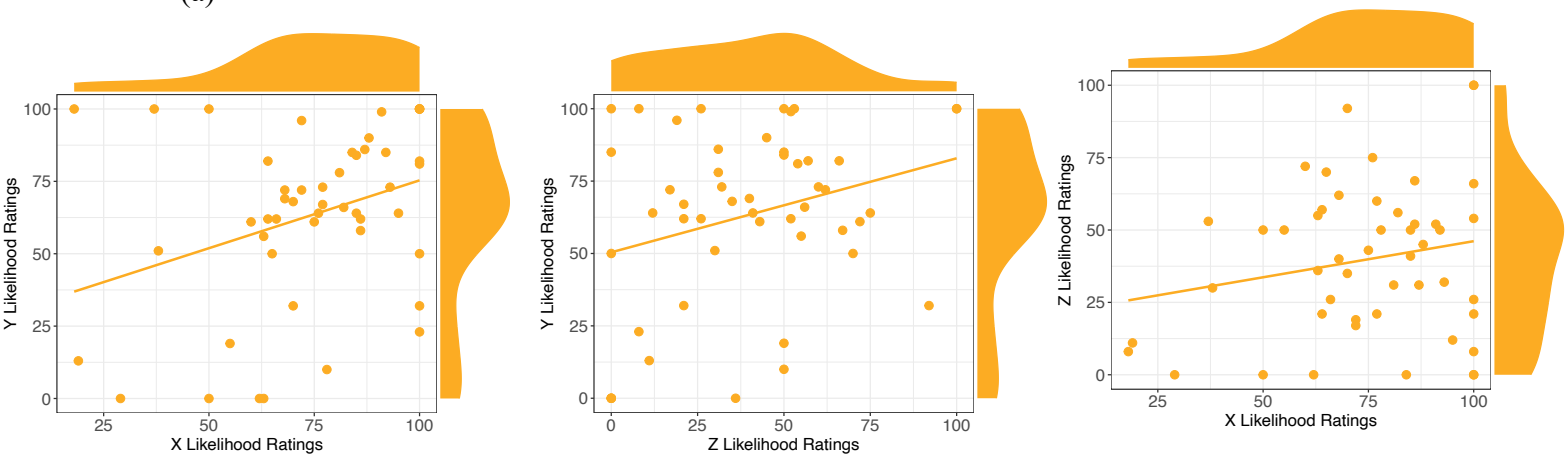


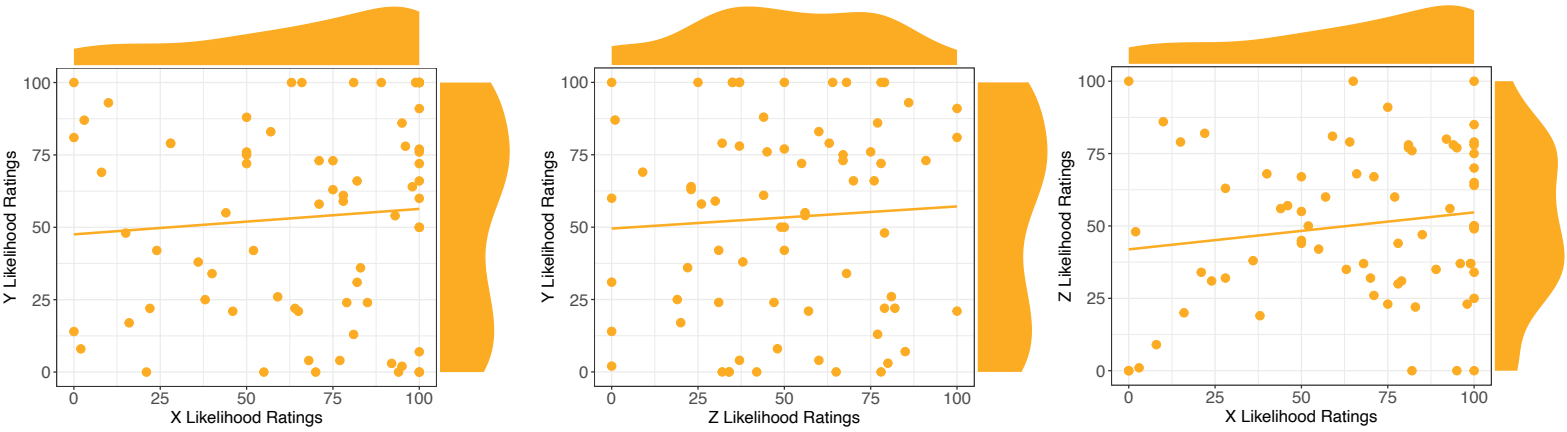
Figure S.6

X vs. Y, Z vs. Y, and X vs. Z correlations simulated under the summed error learning algorithm with the Mackintosh attention across (a) Experiment 2.1, (b) Experiment 2.2, (c) Experiment 2.3, and (d) Experiment 2.4.

(a)



(b)



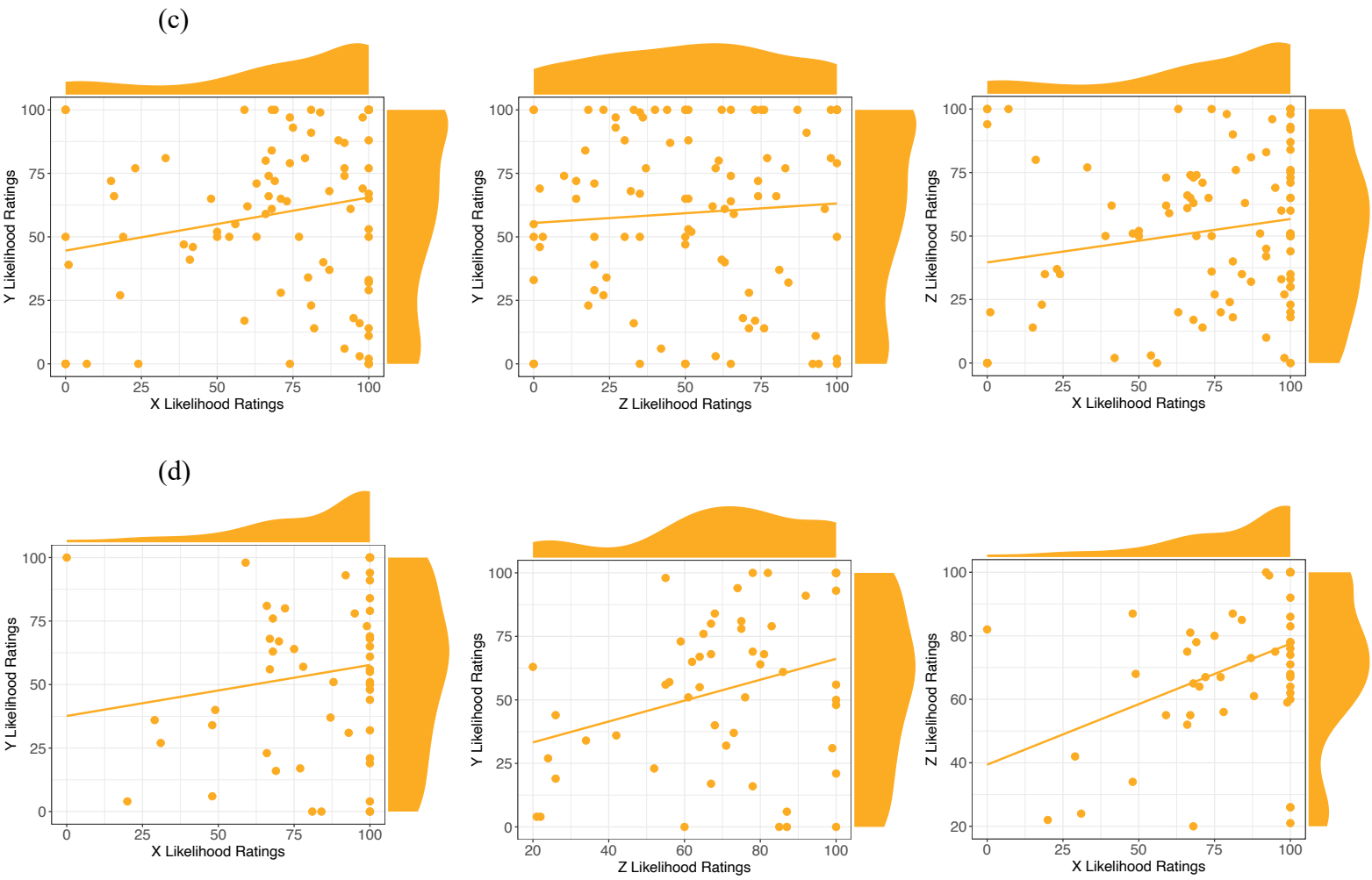
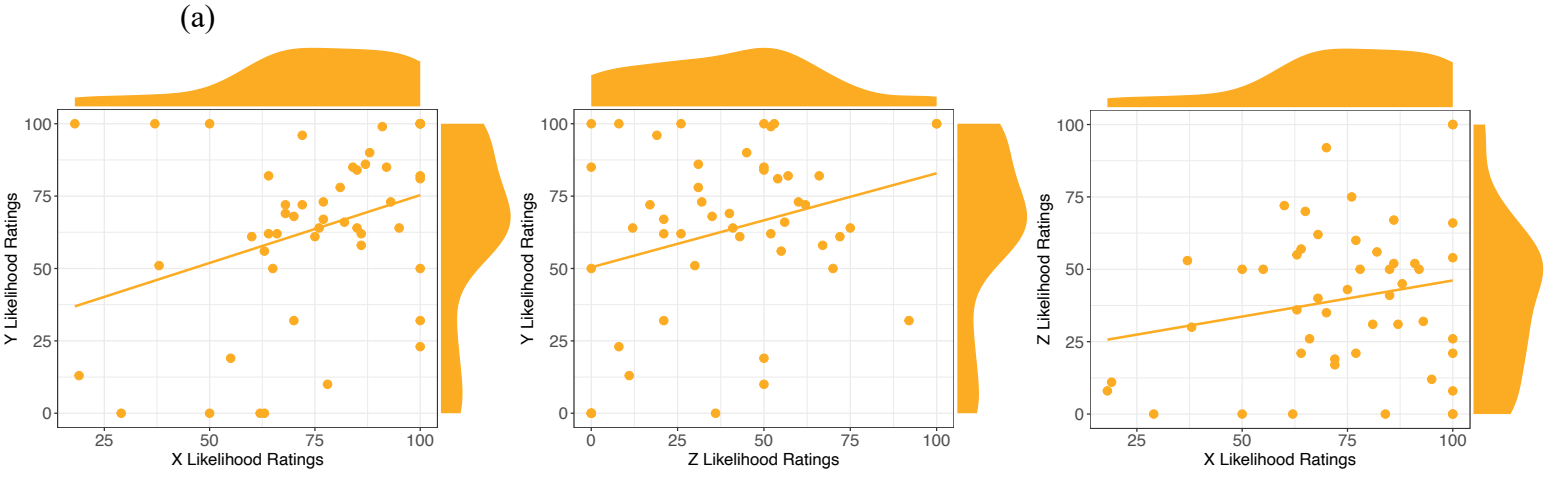
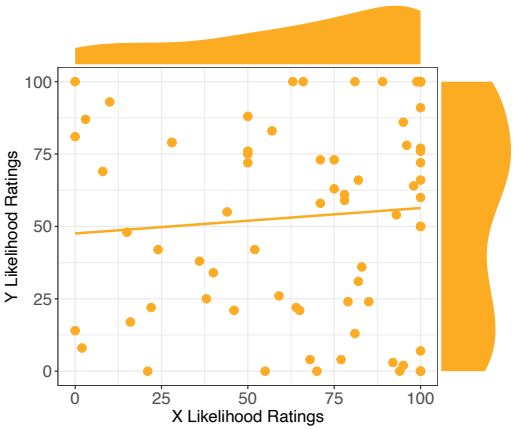


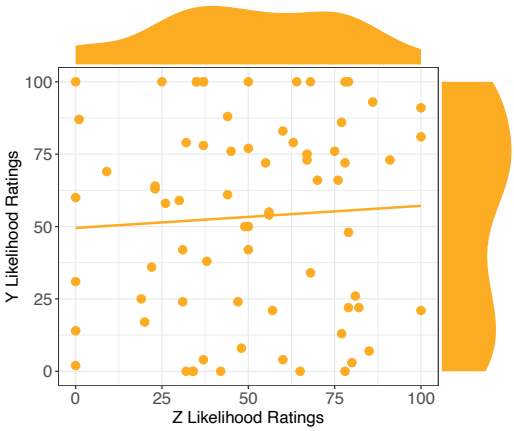
Figure S.7

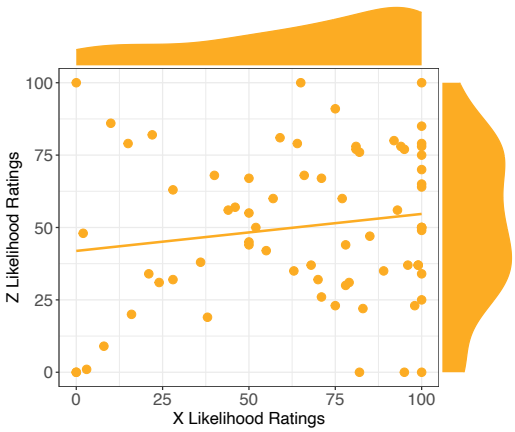
X vs. Y, Z vs. Y, and X vs. Z correlations simulated under the summed error learning algorithm with the Uengoer attention across (a) Experiment 2.1, (b) Experiment 2.2, (c) Experiment 2.3, and (d) Experiment 2.4.



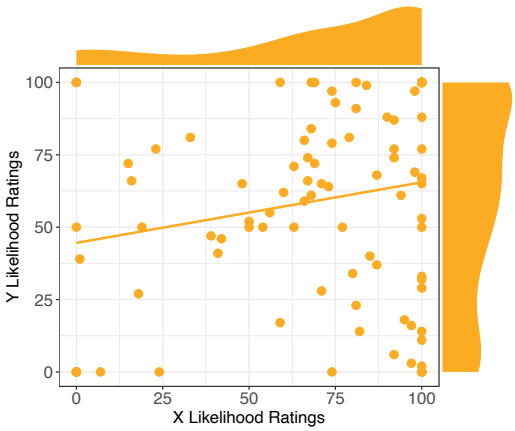
(b)

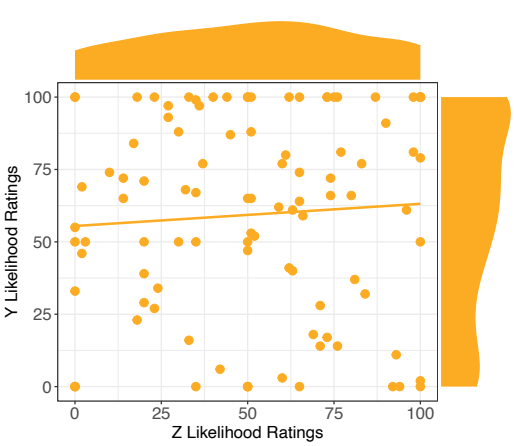

Y Likelihood Ratings
X Likelihood Ratings
0
25
50
75
100

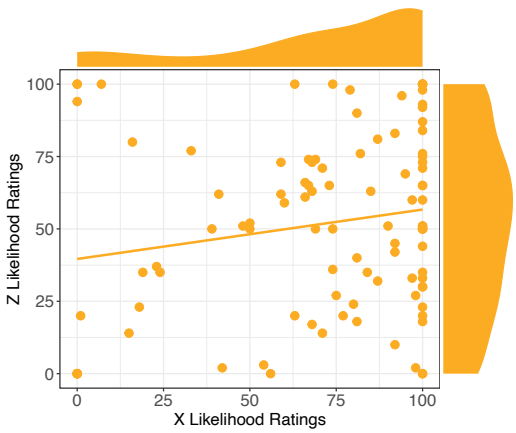

Y Likelihood Ratings
Z Likelihood Ratings
0
25
50
75
100


Z Likelihood Ratings
X Likelihood Ratings
0
25
50
75
100

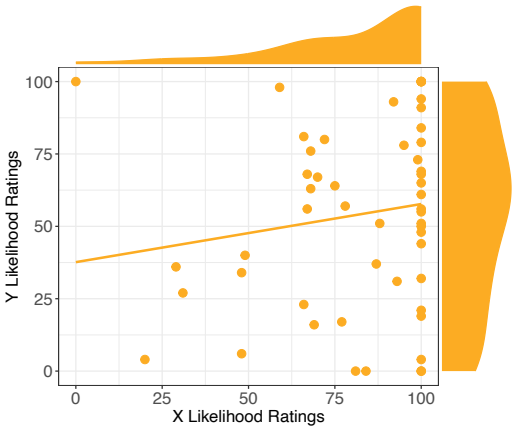
(c)

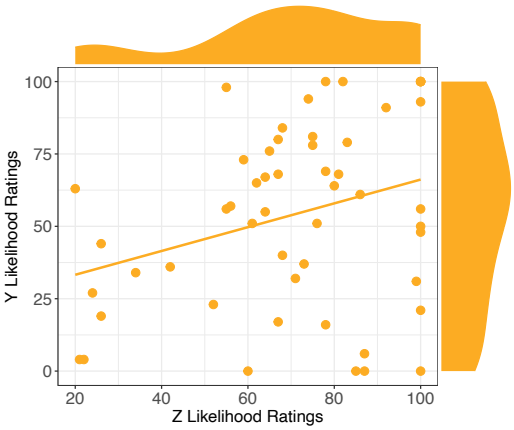

Y Likelihood Ratings
X Likelihood Ratings
0
25
50
75
100

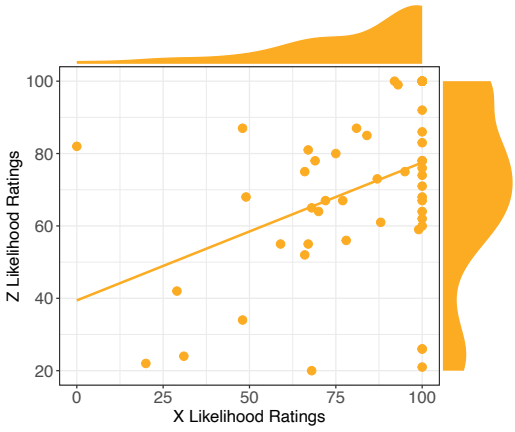

Y Likelihood Ratings
Z Likelihood Ratings
0
25
50
75
100


Z Likelihood Ratings
X Likelihood Ratings
0
25
50
75
100

(d)


Y Likelihood Ratings
X Likelihood Ratings
0
25
50
75
100


Y Likelihood Ratings
Z Likelihood Ratings
20
40
60
80
100


Z Likelihood Ratings
X Likelihood Ratings
0
25
50
75
100
20
40
60
80

Simulated Associability Change

Figure S.8

Associability change across training simulated under the separable error learning algorithm with the Mackintosh attention for (a) Experiment 2.1, (b) Experiment 2.2, (c) the 0% group of Experiment 2.3, (d) the 50% group of Experiment 2.3, and (e) Experiment 2.4.

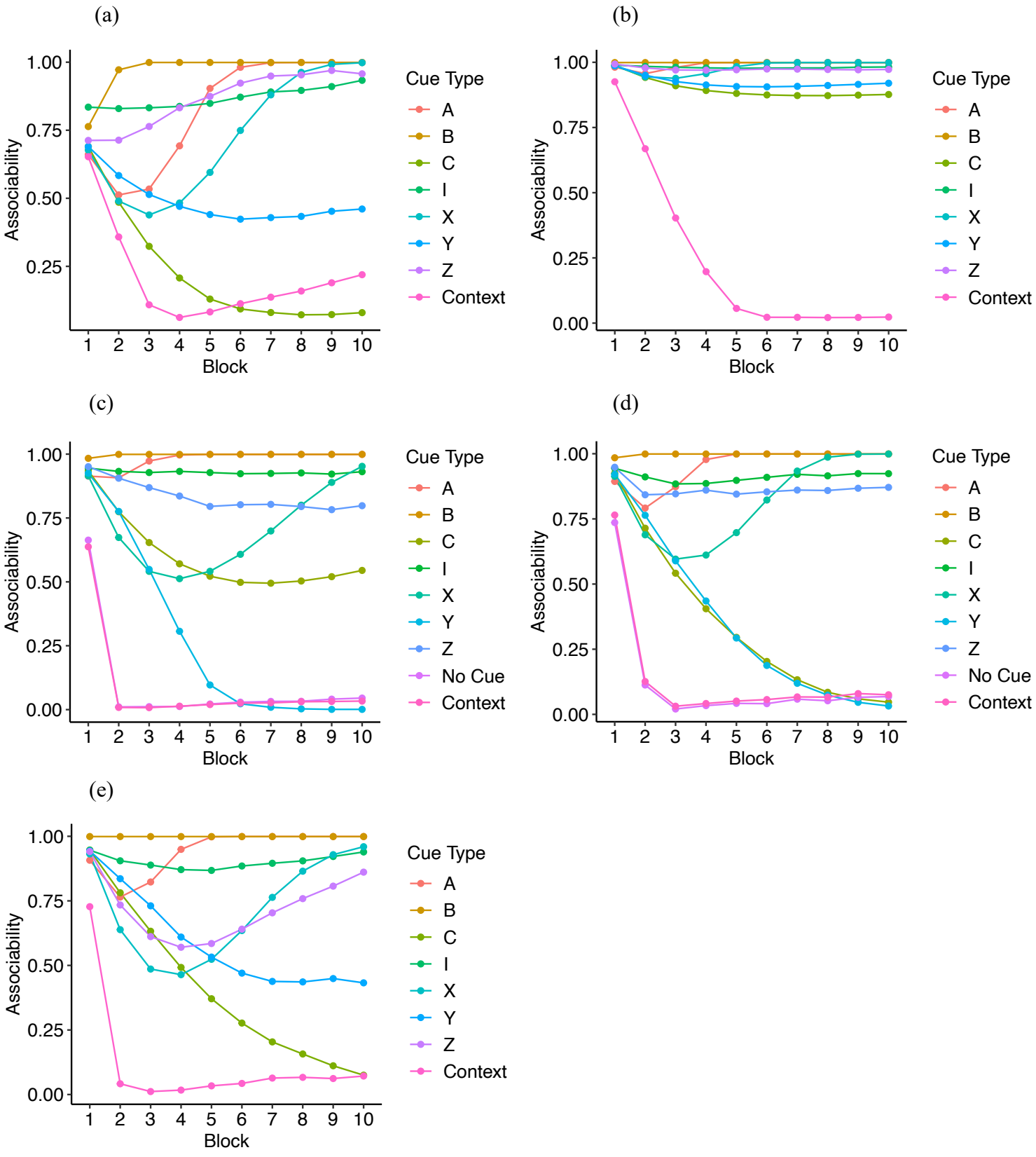


Figure S.9

Associability change across training simulated under the summed error learning algorithm with the Mackintosh attention for (a) Experiment 2.1, (b) Experiment 2.2, (c) the 0% group of Experiment 2.3, (d) the 50% group of Experiment 2.3, and (e) Experiment 2.4.

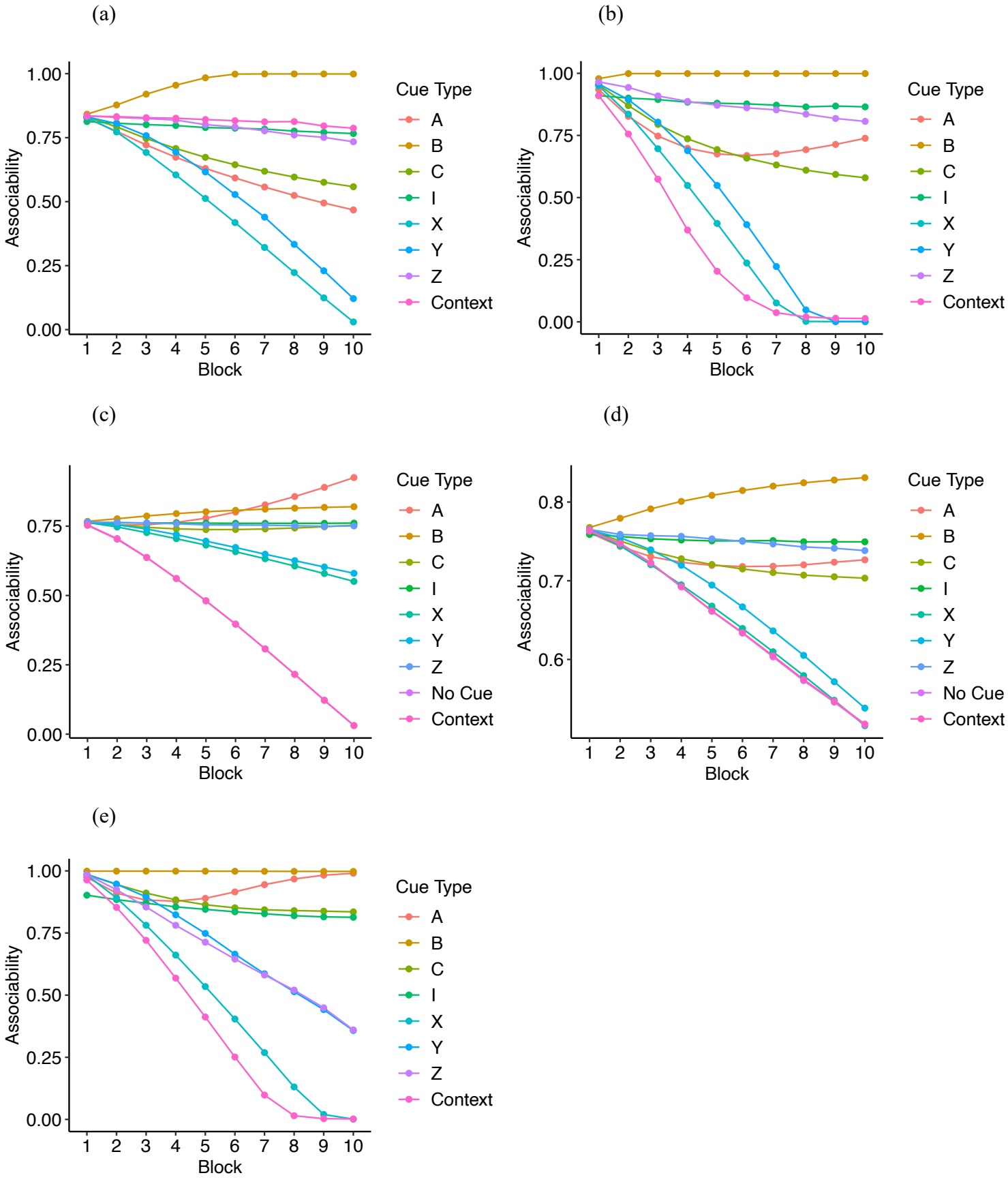


Figure S.10

Associability change across training simulated under the separable error learning algorithm with the Uengoer attention for (a) Experiment 2.1, (b) Experiment 2.2, (c) the 0% group of Experiment 2.3, (d) the 50% group of Experiment 2.3, and (e) Experiment 2.4.

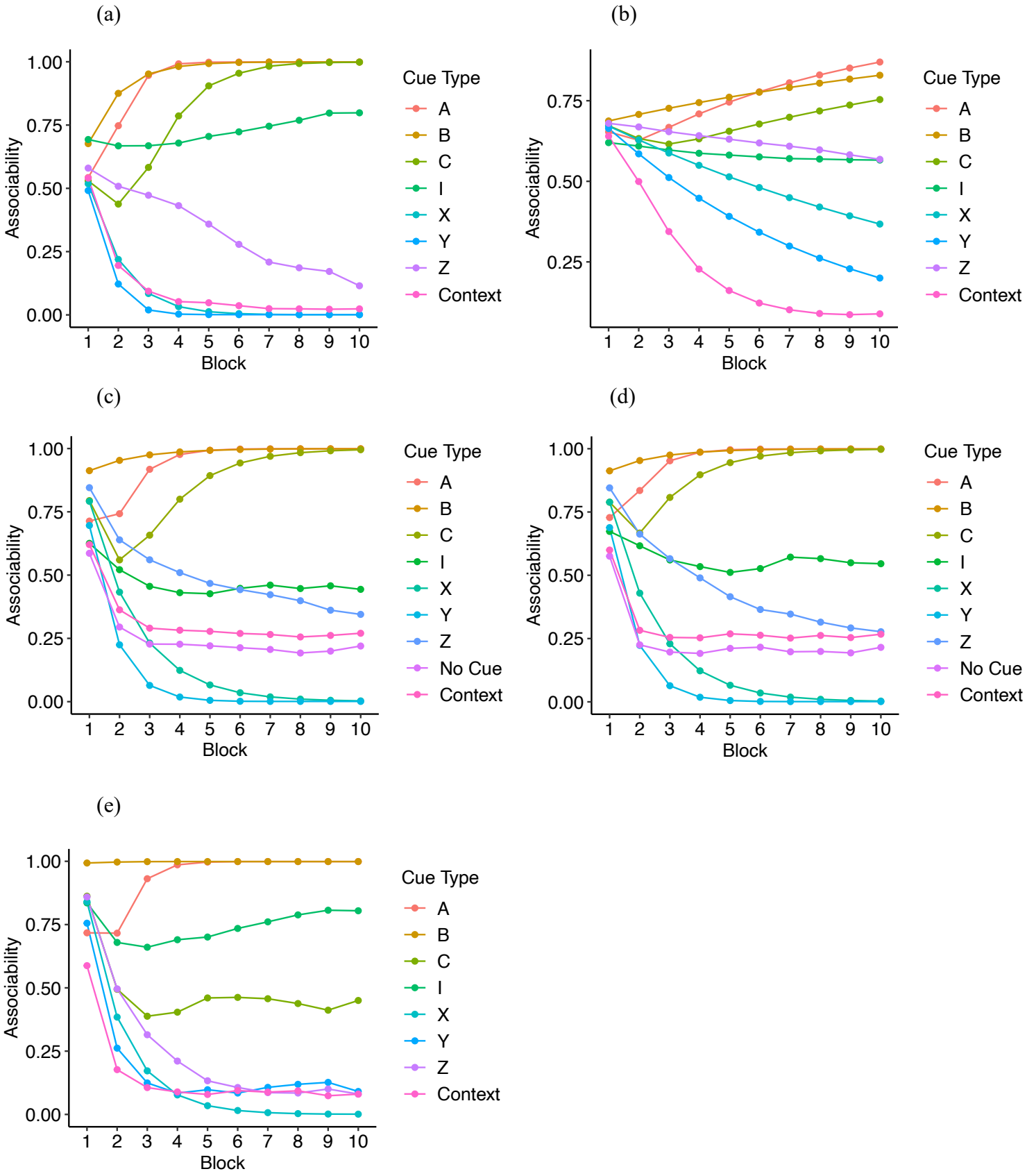
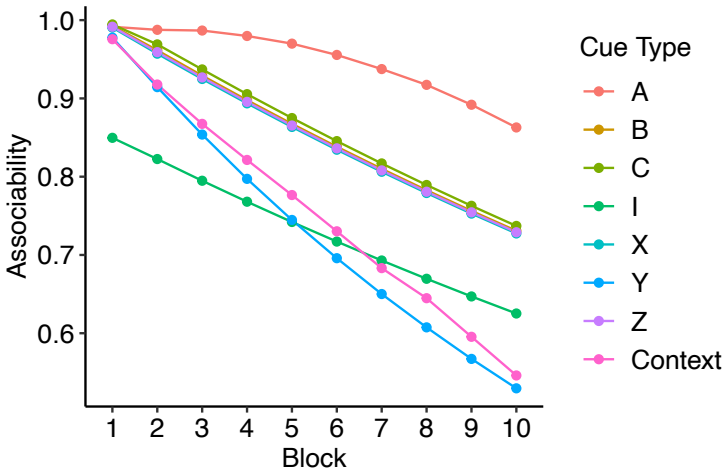


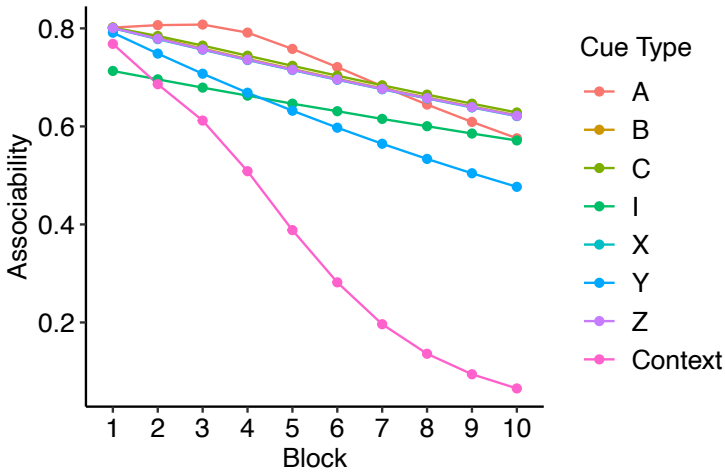
Figure S.11

Associability change across training simulated under the summed error learning algorithm with the Uengoer attention for (a) Experiment 2.1, (b) Experiment 2.2, (c) the 0% group of Experiment 2.3, (d) the 50% group of Experiment 2.3, and (e) Experiment 2.4.

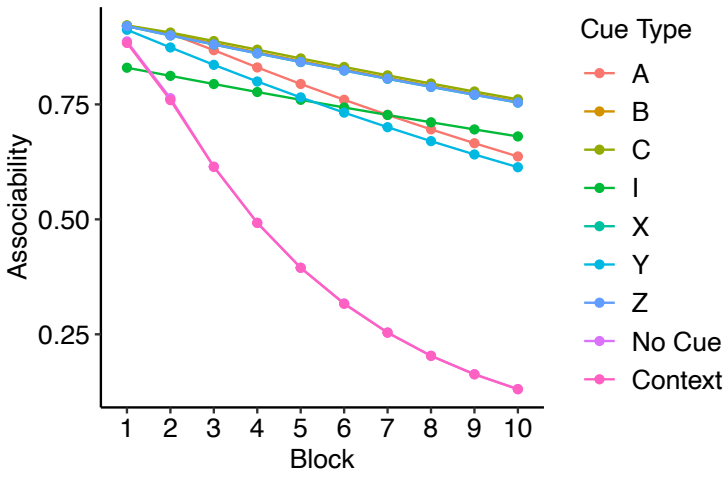
(a)



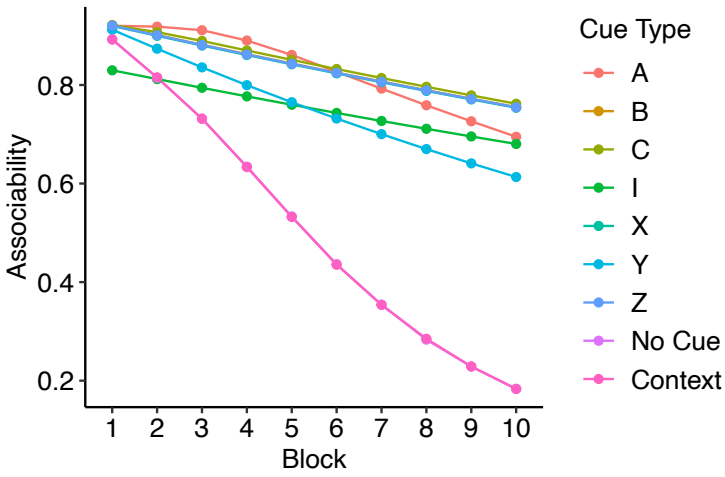
(b)



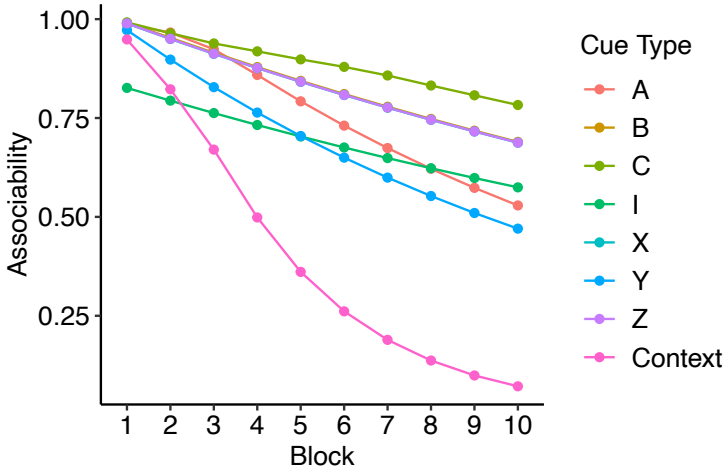
(c)



(d)



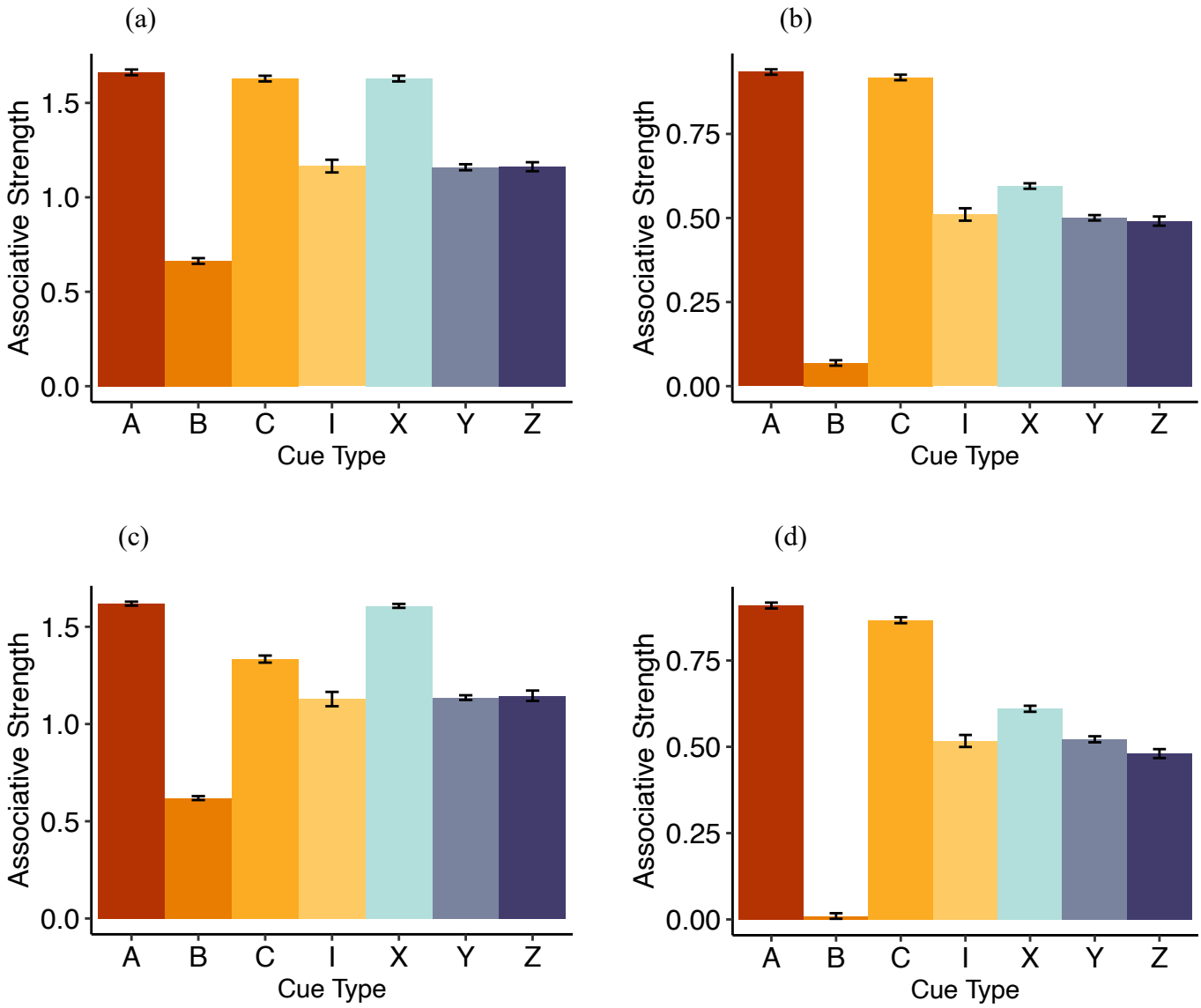
(e)



Simulated Likelihood Ratings

Figure S.12

Simulated associative strengths for cues in the likelihood ratings test of Experiment 2.1 under (a) the separable error term model without attention, (b) the summed error term model without attention, (c) the separable error term model with the Mackintosh attention, (d) the summed error model with the Mackintosh attention, (e) the separable error term model with the Uengoer attention, and (f) the summed error term model with the Uengoer attention.



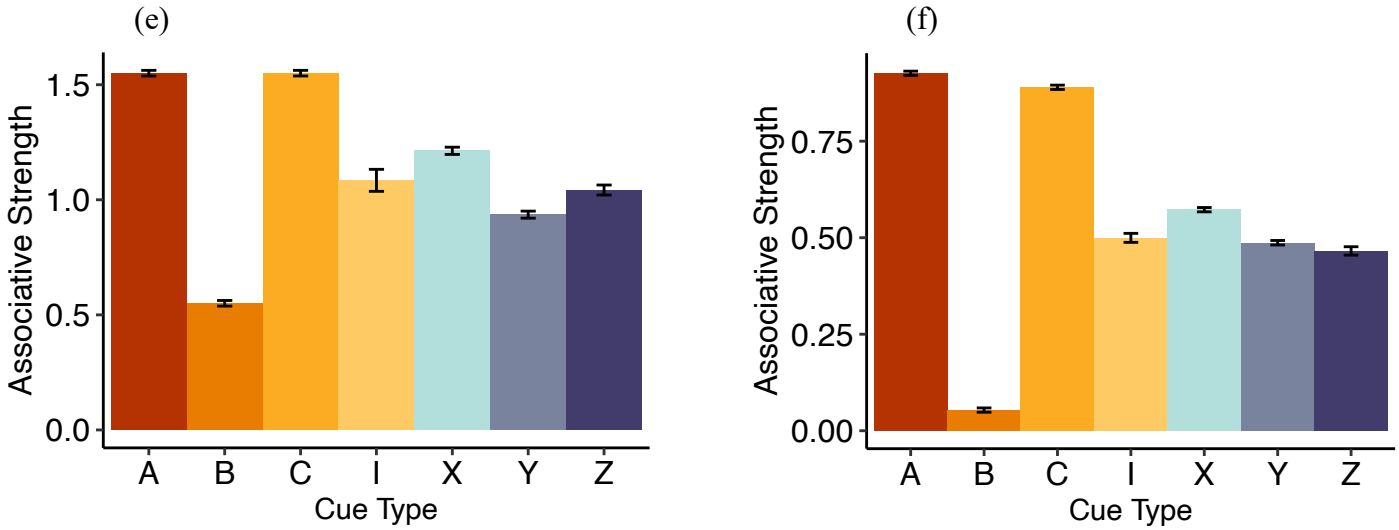
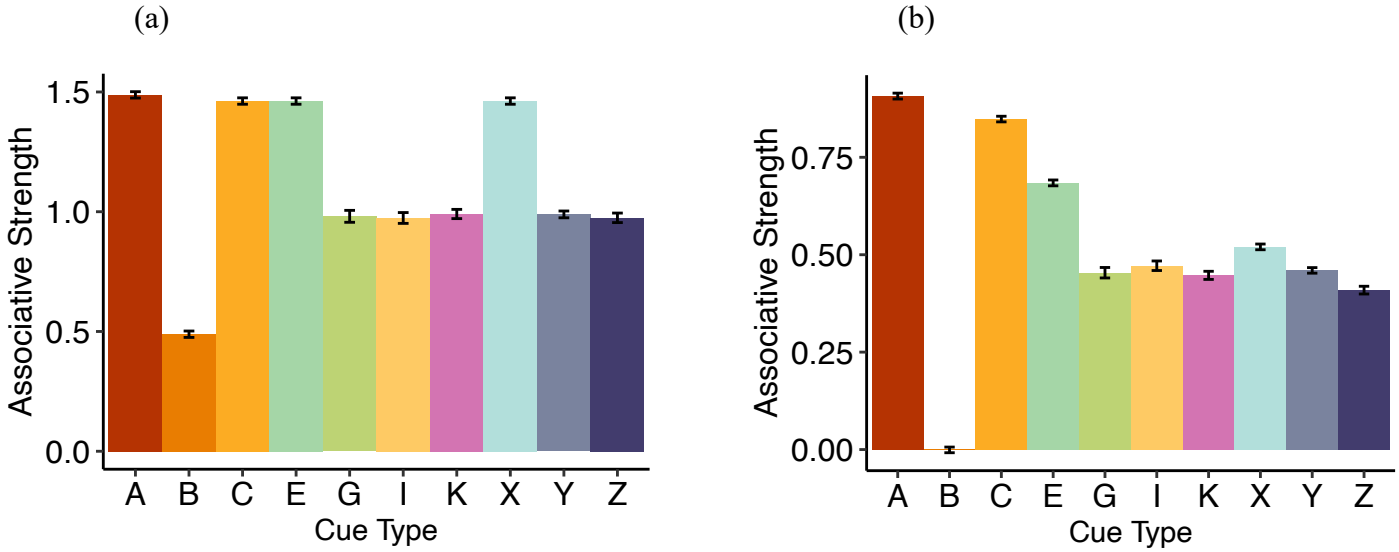


Figure S.13

Simulated associative strengths for cues in the likelihood ratings test of Experiment 2.2 under (a) the separable error term model without attention, (b) the summed error term model without attention, (c) the separable error term model with the Mackintosh attention, (d) the summed error model with the Mackintosh attention, (e) the separable error term model with the Uengoer attention, and (f) the summed error term model with the Uengoer attention.



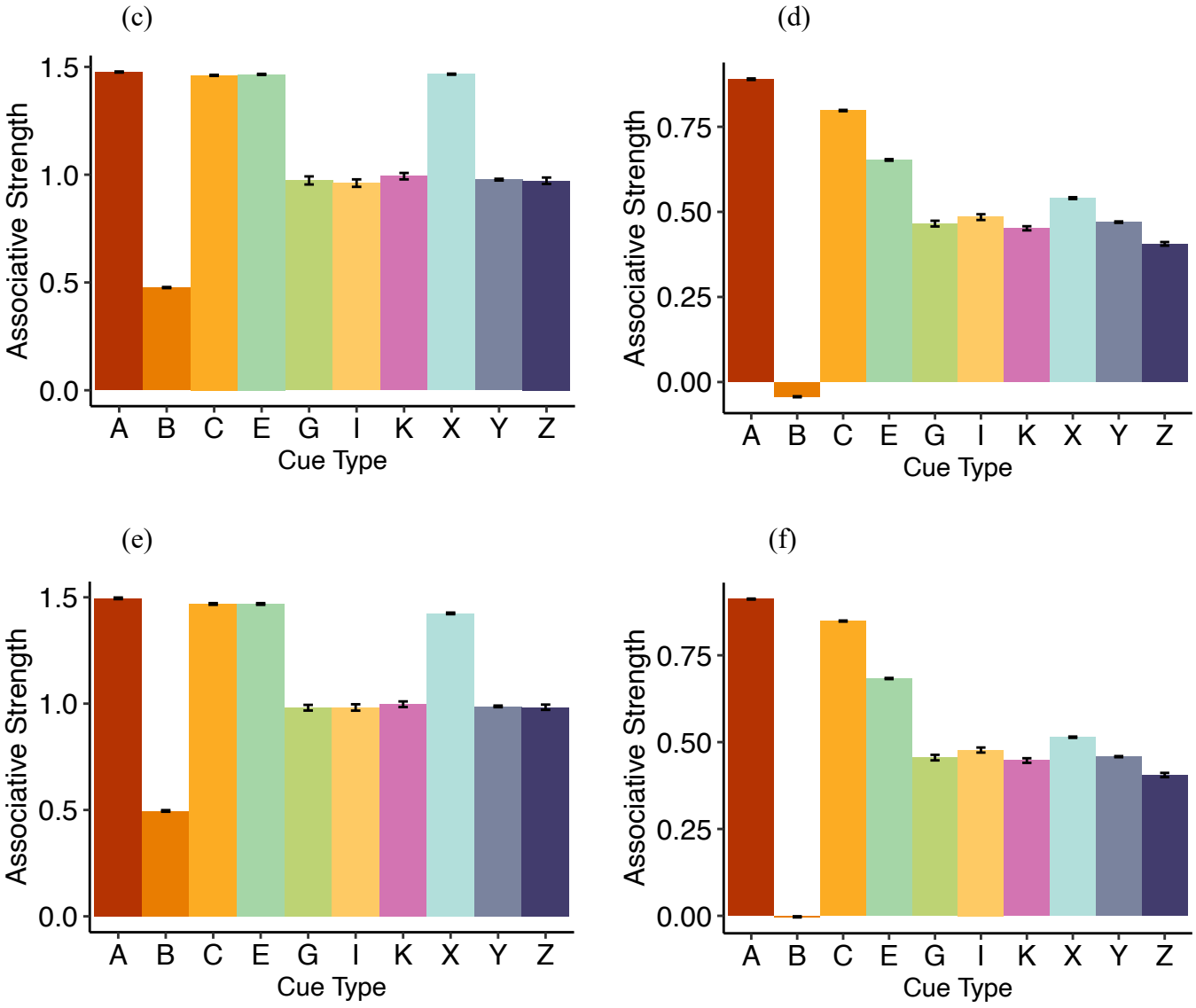
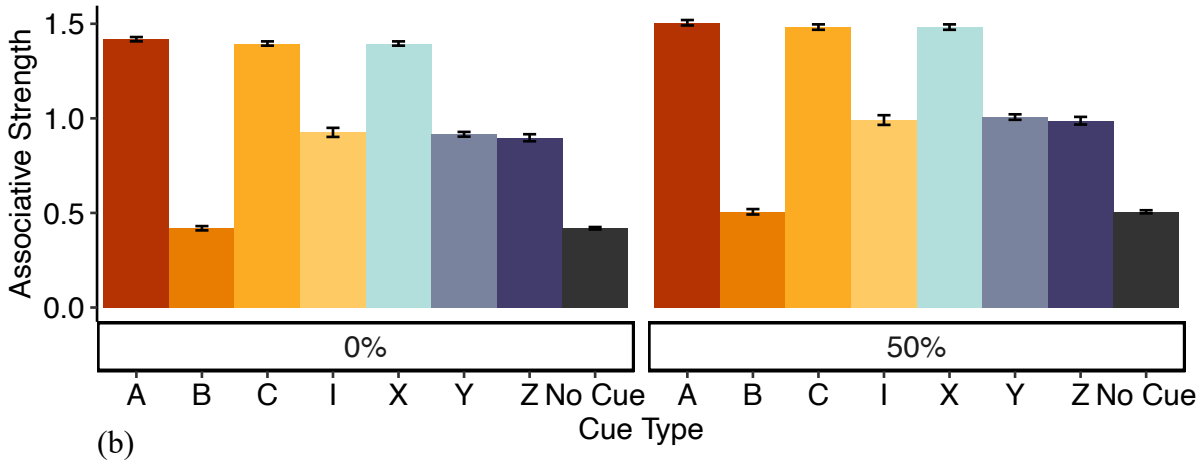


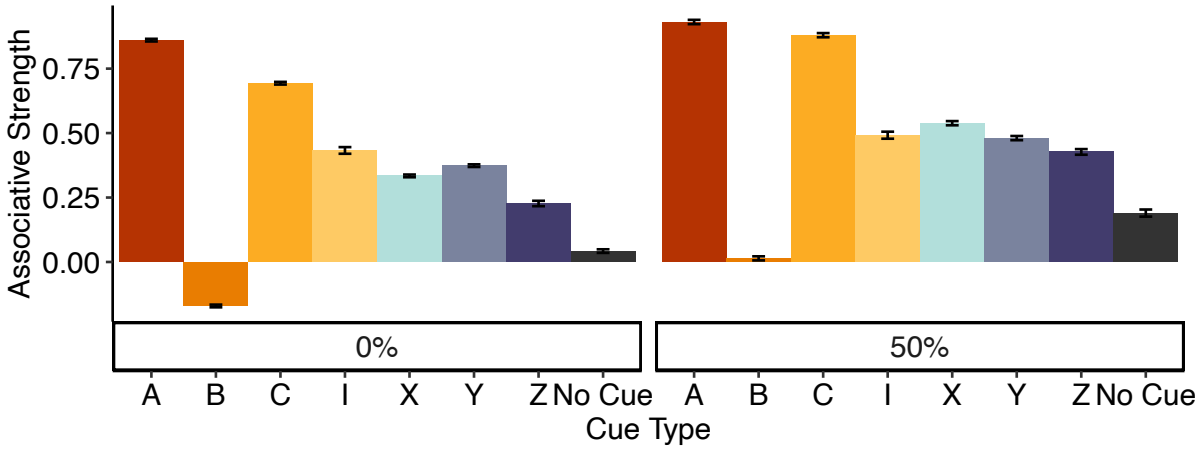
Figure S.14

Simulated associative strengths for cues in the likelihood ratings test of Experiment 2.3 under (a) the separable error term model without attention, (b) the summed error term model without attention, (c) the separable error term model with the Mackintosh attention, (d) the summed error model with the Mackintosh attention, (e) the separable error term model with the Uengoer attention, and (f) the summed error term model with the Uengoer attention.

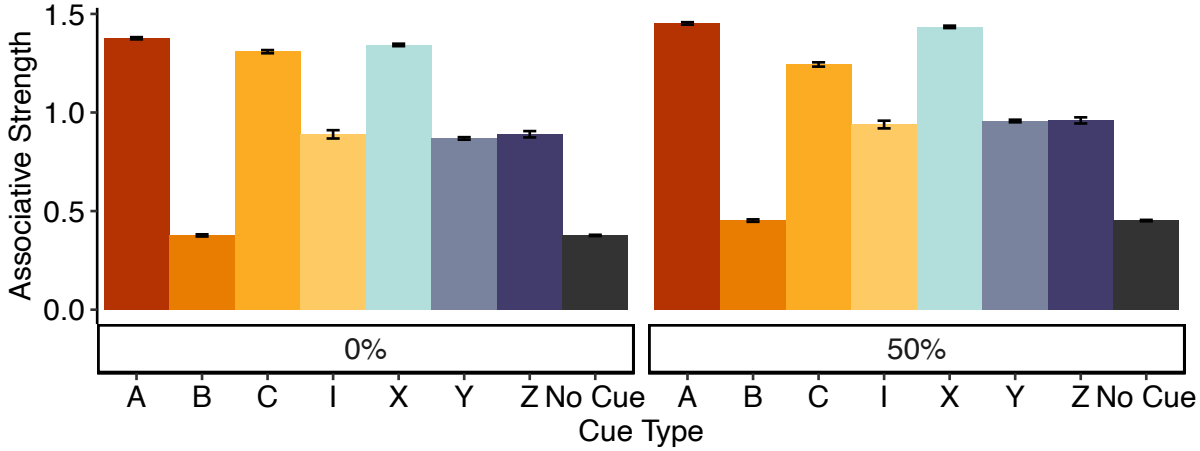
(a)


Associative Strength
1.5
1.0
0.5
0.0
0%
50%
A B C I X Y Z No Cue A B C I X Y Z No Cue
Cue Type

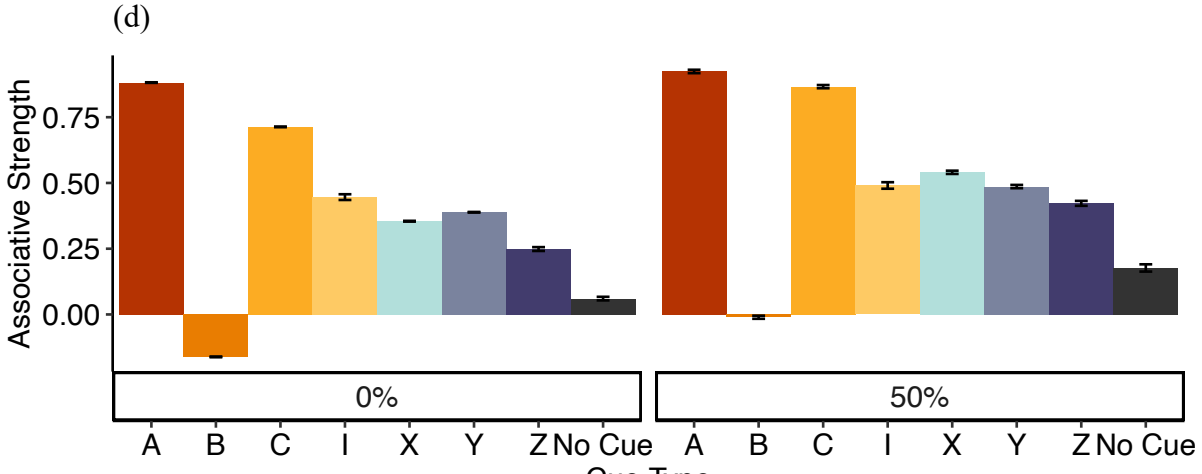
(b)


Associative Strength
0.75
0.50
0.25
0.00
0%
50%
A B C I X Y Z No Cue A B C I X Y Z No Cue
Cue Type

(c)


Associative Strength
1.5
1.0
0.5
0.0
0%
50%
A B C I X Y Z No Cue A B C I X Y Z No Cue
Cue Type

(d)


Associative Strength
0.75
0.50
0.25
0.00
0%
50%
A B C I X Y Z No Cue A B C I X Y Z No Cue
Cue Type

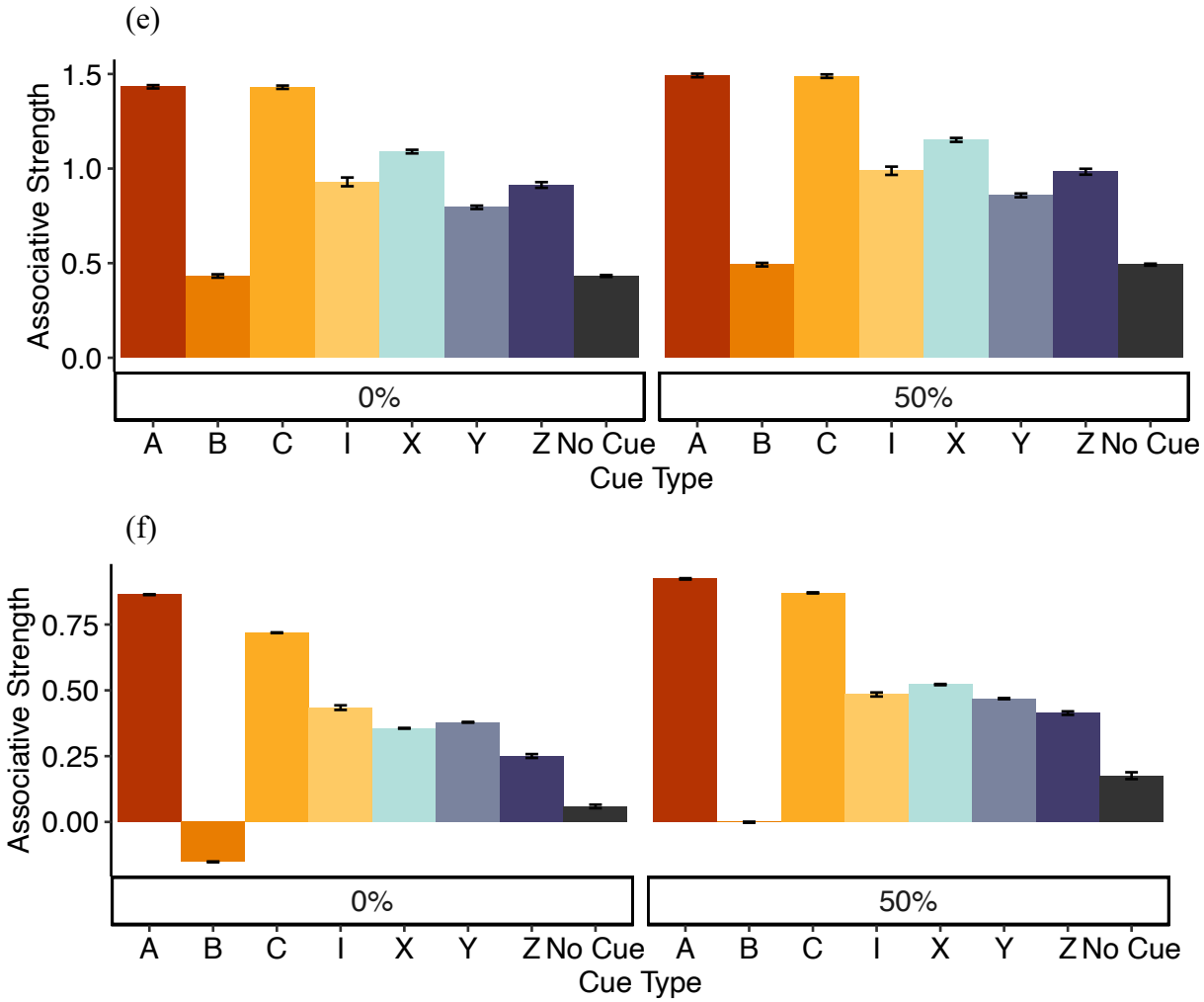
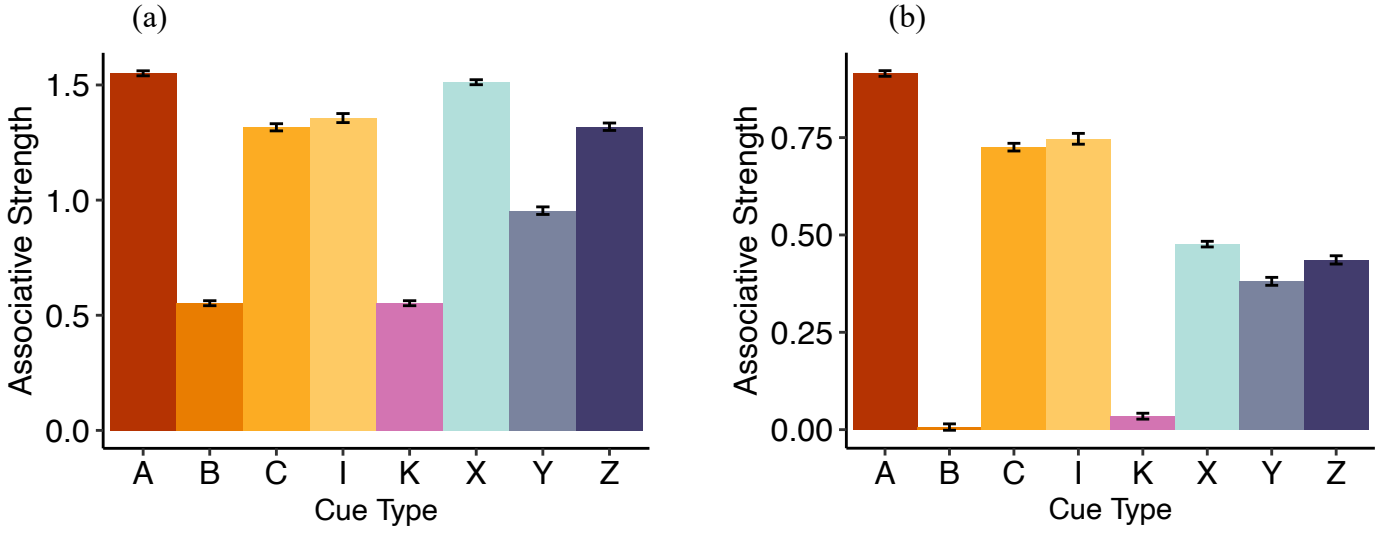
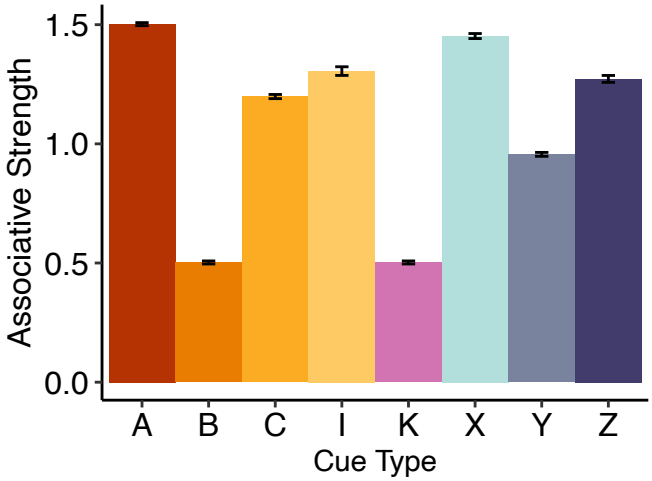


Figure S.15

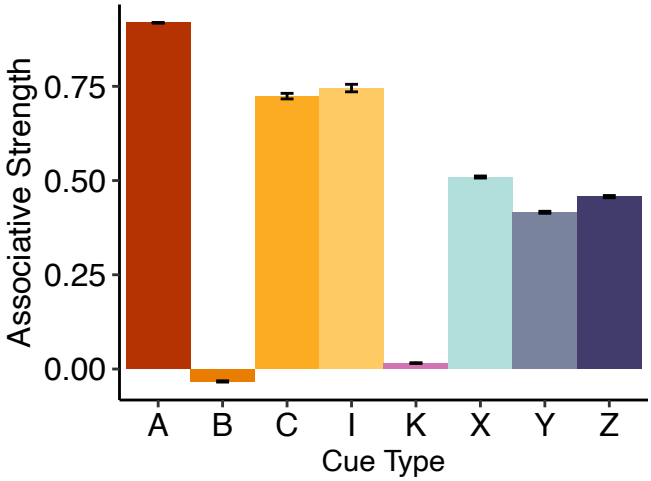
Simulated associative strengths for cues in the likelihood ratings test of Experiment 2.4 under (a) the separable error term model without attention, (b) the summed error term model without attention, (c) the separable error term model with the Mackintosh attention, (d) the summed error model with the Mackintosh attention, (e) the separable error term model with the Uengoer attention, and (f) the summed error term model with the Uengoer attention.



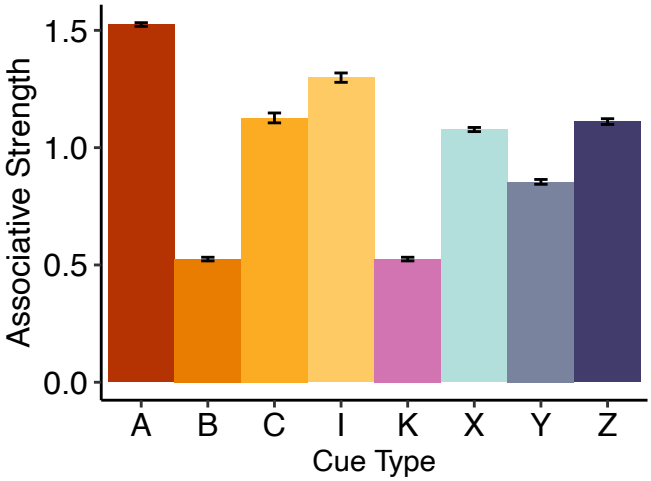
(c)


Associative Strength
1.5
1.0
0.5
0.0
A B C I K X Y Z
Cue Type

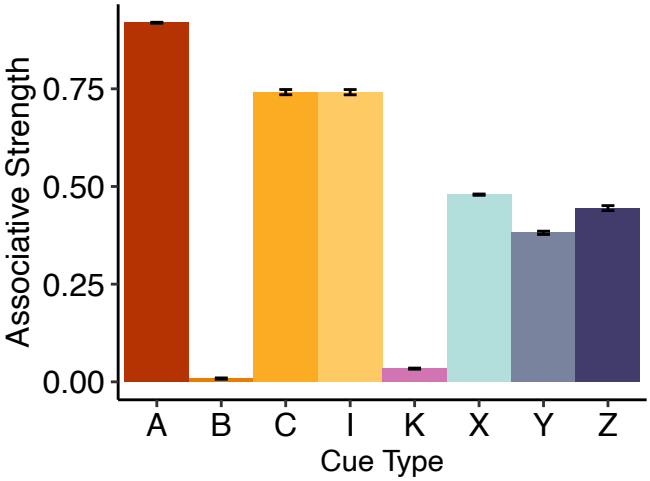
(d)


Associative Strength
0.75
0.50
0.25
0.00
A B C I K X Y Z
Cue Type

(e)


Associative Strength
1.5
1.0
0.5
0.0
A B C I K X Y Z
Cue Type

(f)

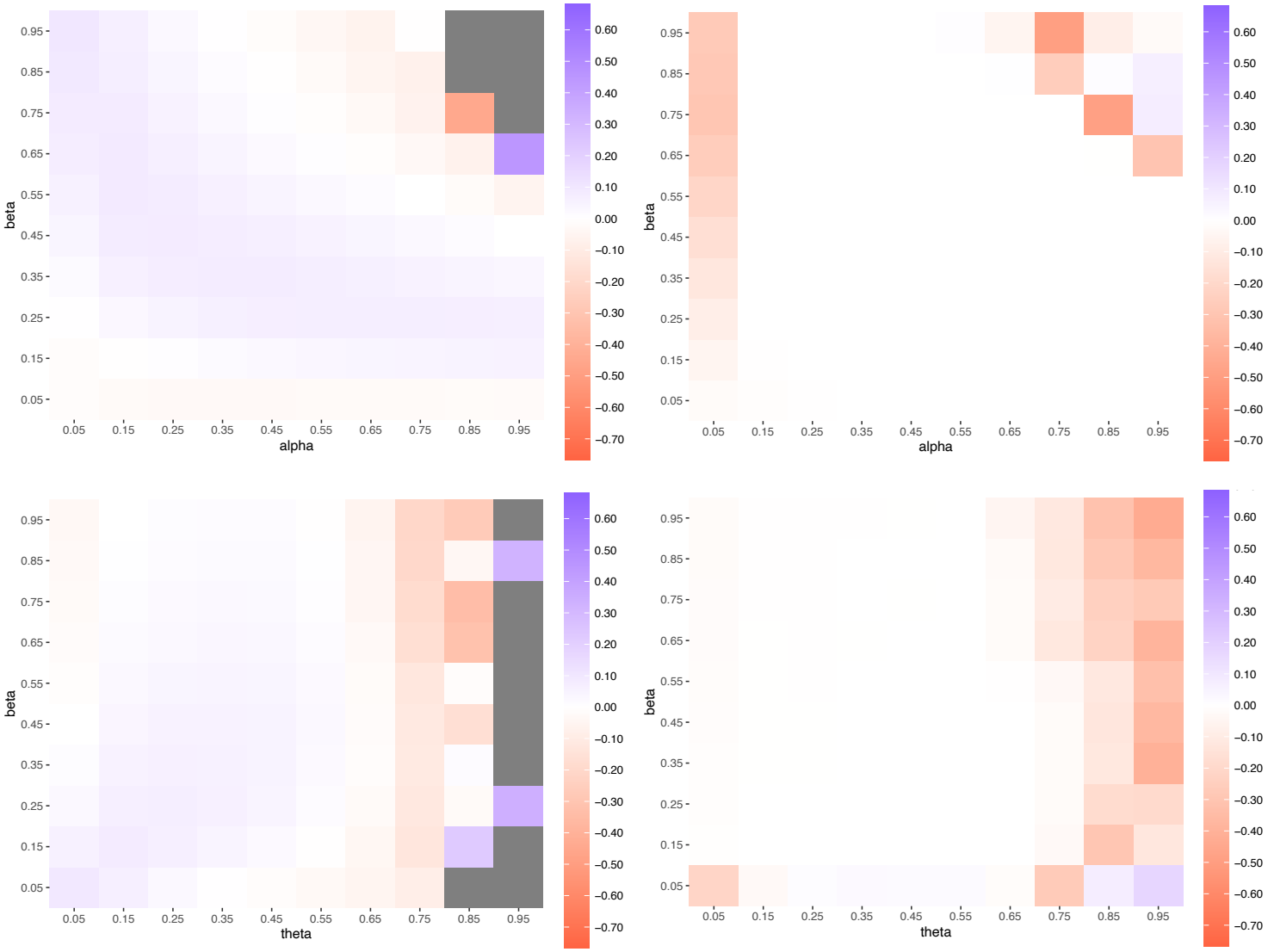

Associative Strength
0.75
0.50
0.25
0.00
A B C I K X Y Z
Cue Type

Grid Search

Figure S.16

Differences in associative strength (left panels) and associability (right panels) across different alpha-beta and theta-beta combinations for (a) XY and (b) ZY simulated under the summed error learning algorithm with the Mackintosh attention in Experiment 2.1.

(a)



(b)

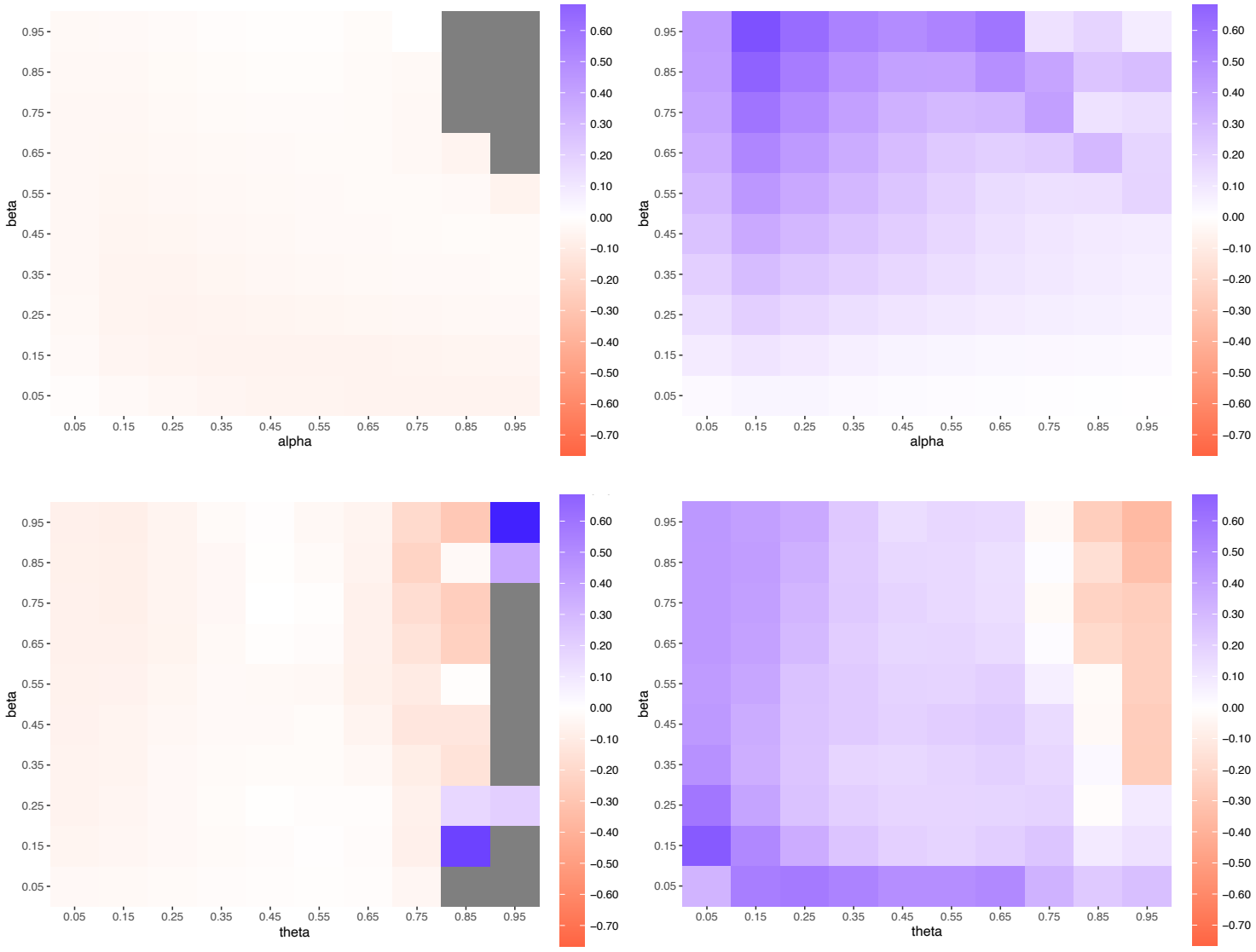
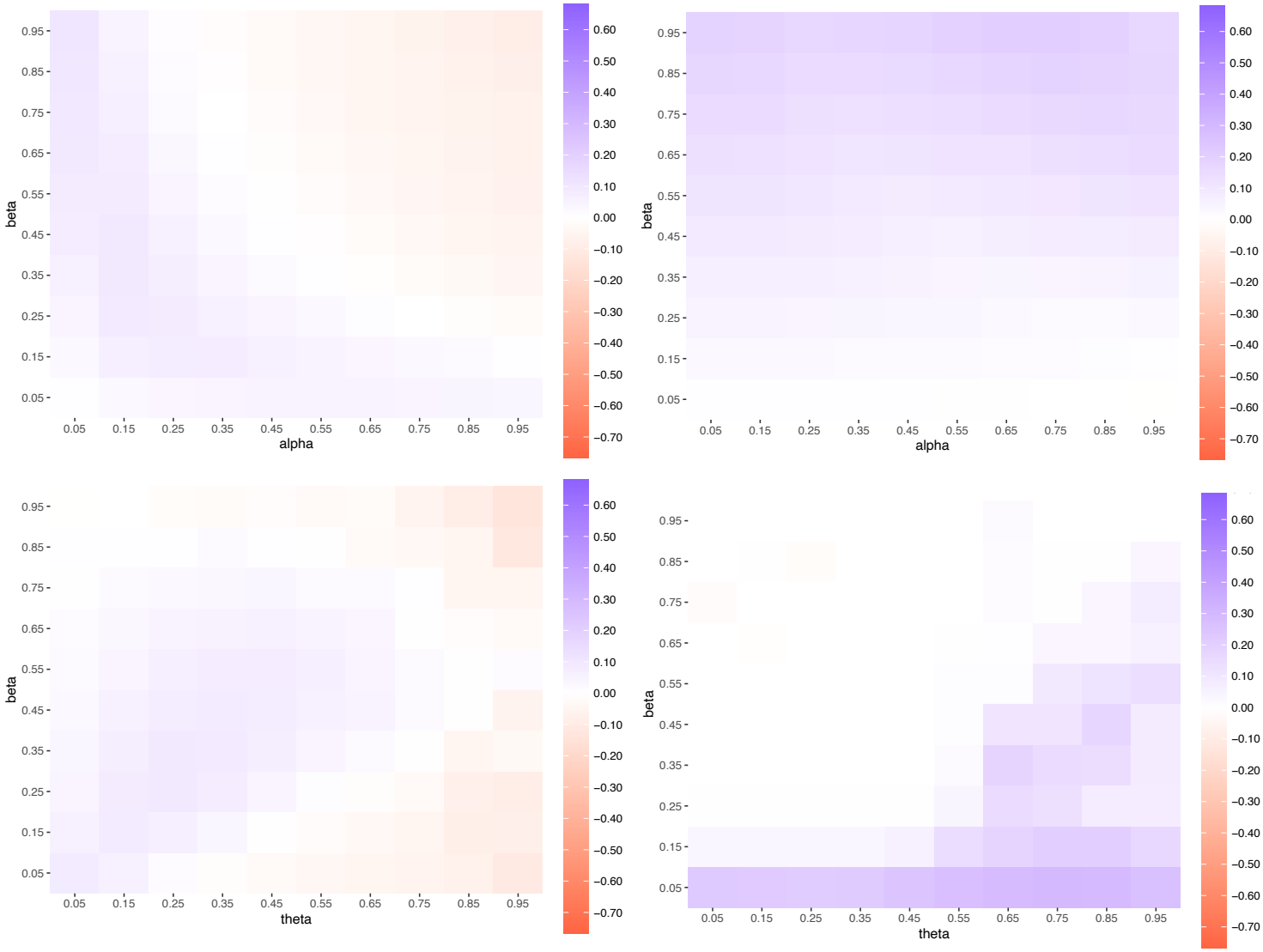

beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

Figure S.17

Differences in associative strength (left panels) and associability (right panels) across different alpha-beta and theta-beta combinations for (a) XY and (b) ZY simulated under the summed error learning algorithm with the Uengeor attention in Experiment 2.1.

(a)



(b)

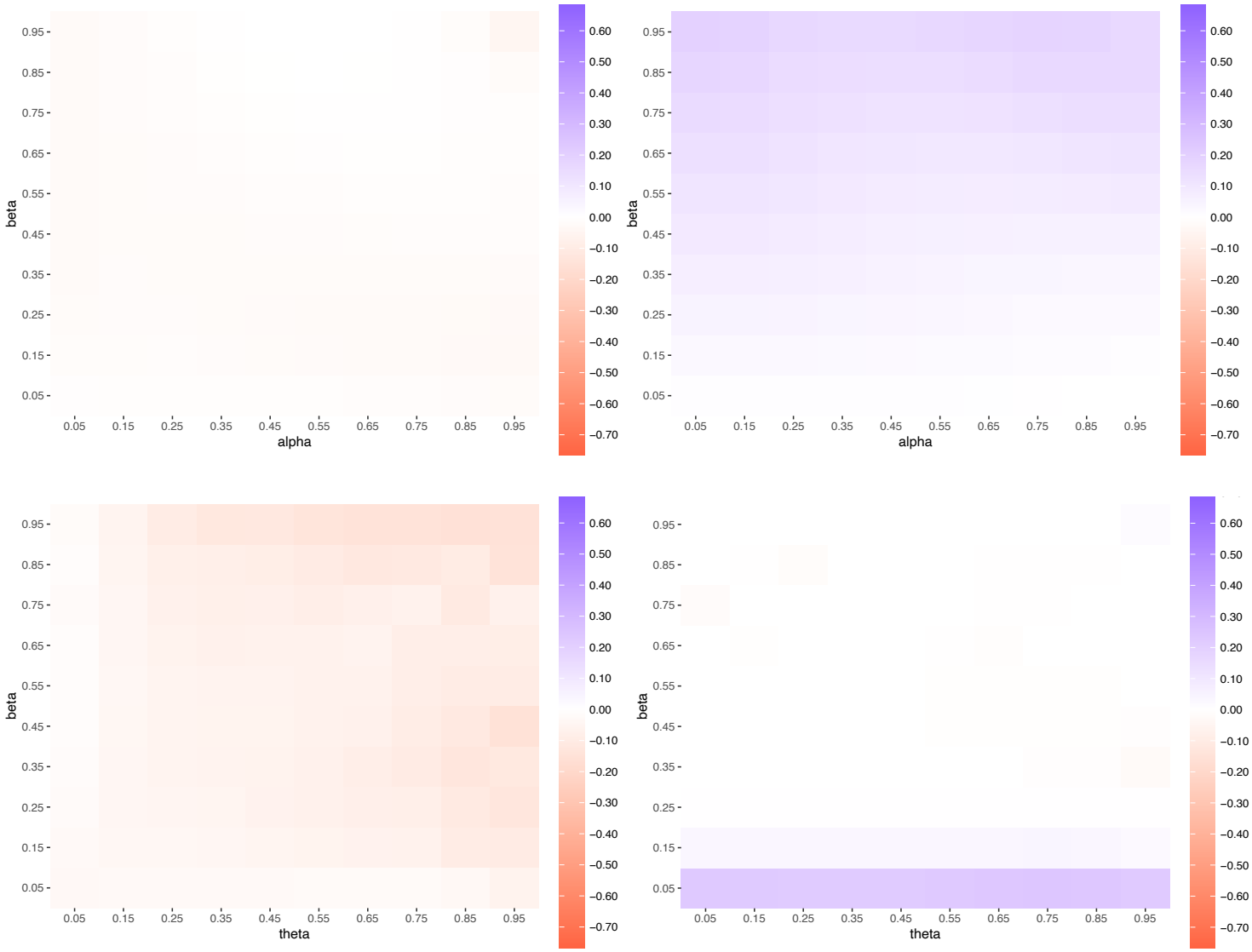
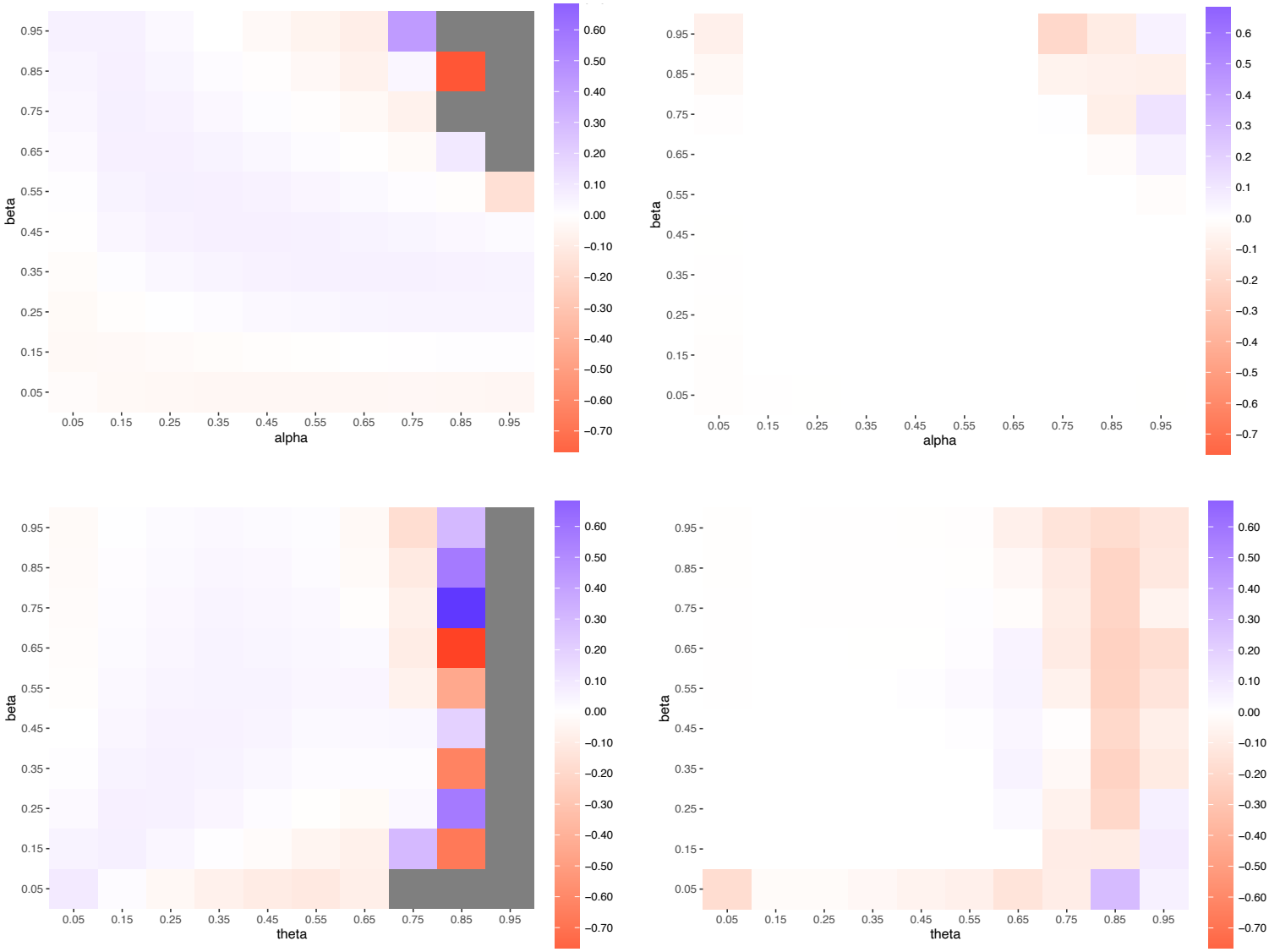

beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

Figure S.18

Differences in associative strength (left panels) and associability (right panels) across different alpha-beta and theta-beta combinations for (a) XY and (b) ZY simulated under the summed error learning algorithm with the Mackintosh attention in Experiment 2.2.

(a)



(b)

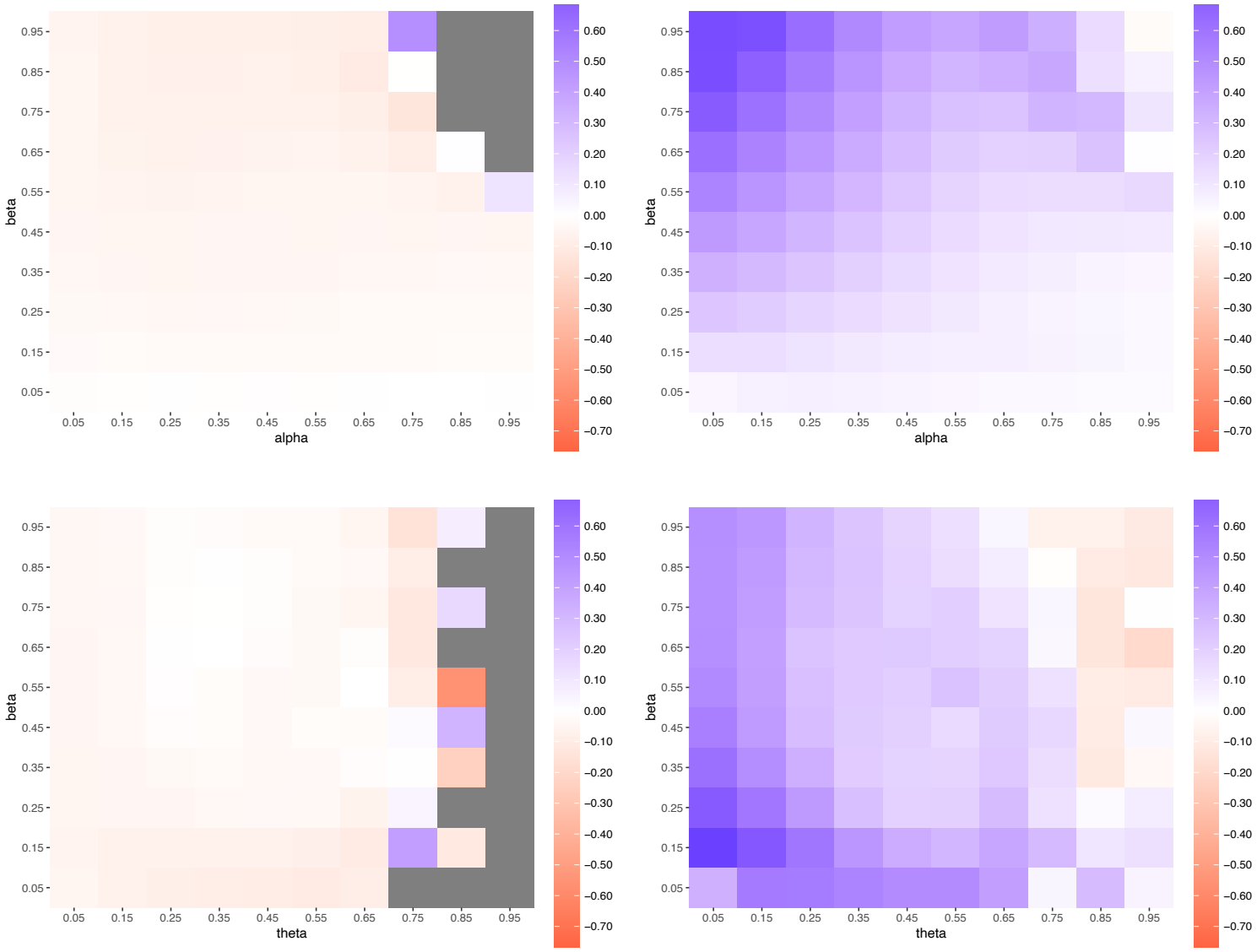
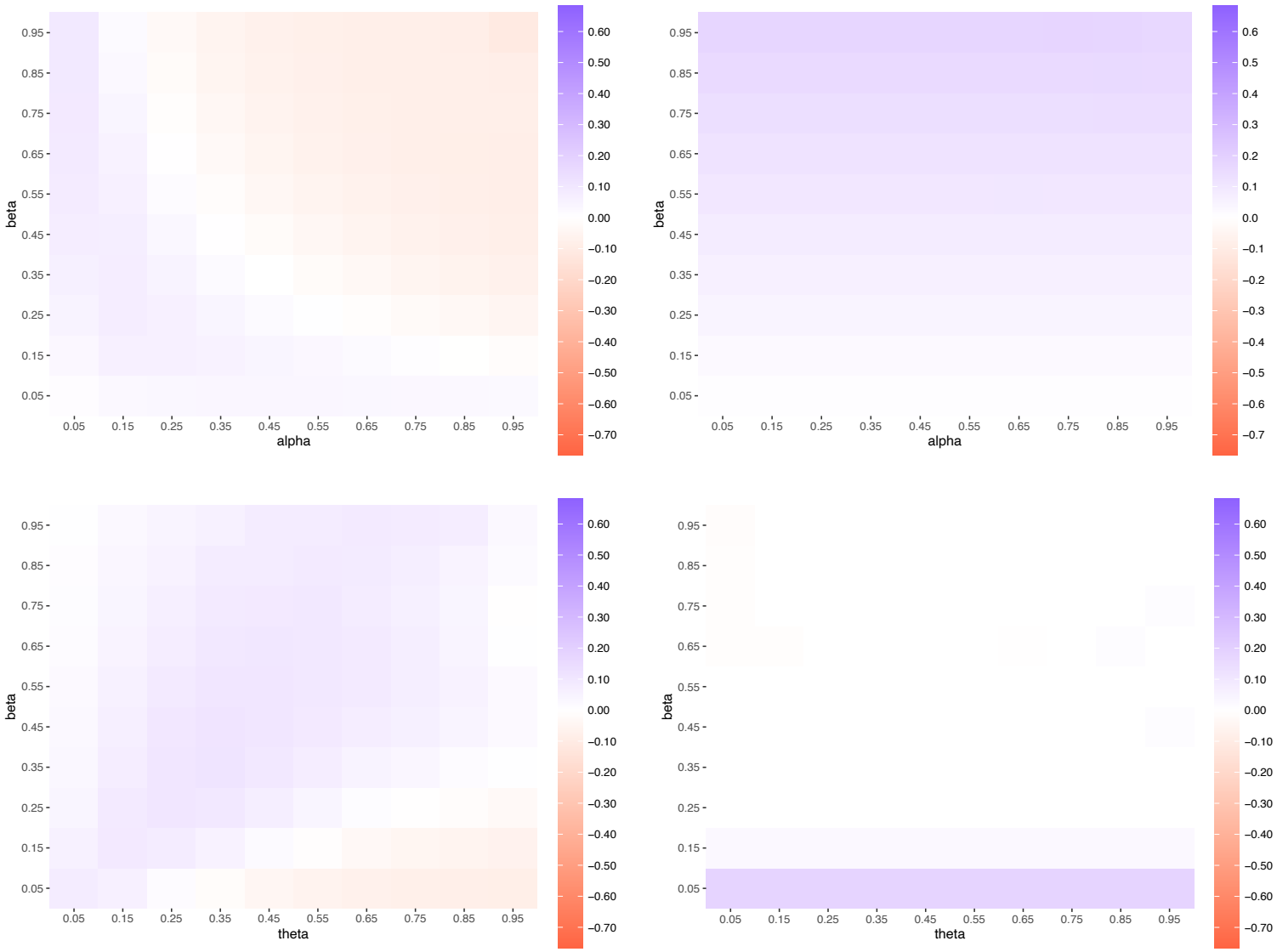

beta
alpha
theta
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

Figure S.19

Differences in associative strength (left panels) and associability (right panels) across different alpha-beta and theta-beta combinations for (a) XY and (b) ZY simulated under the summed error learning algorithm with the Uengoer attention in Experiment 2.2.

(a)



(b)

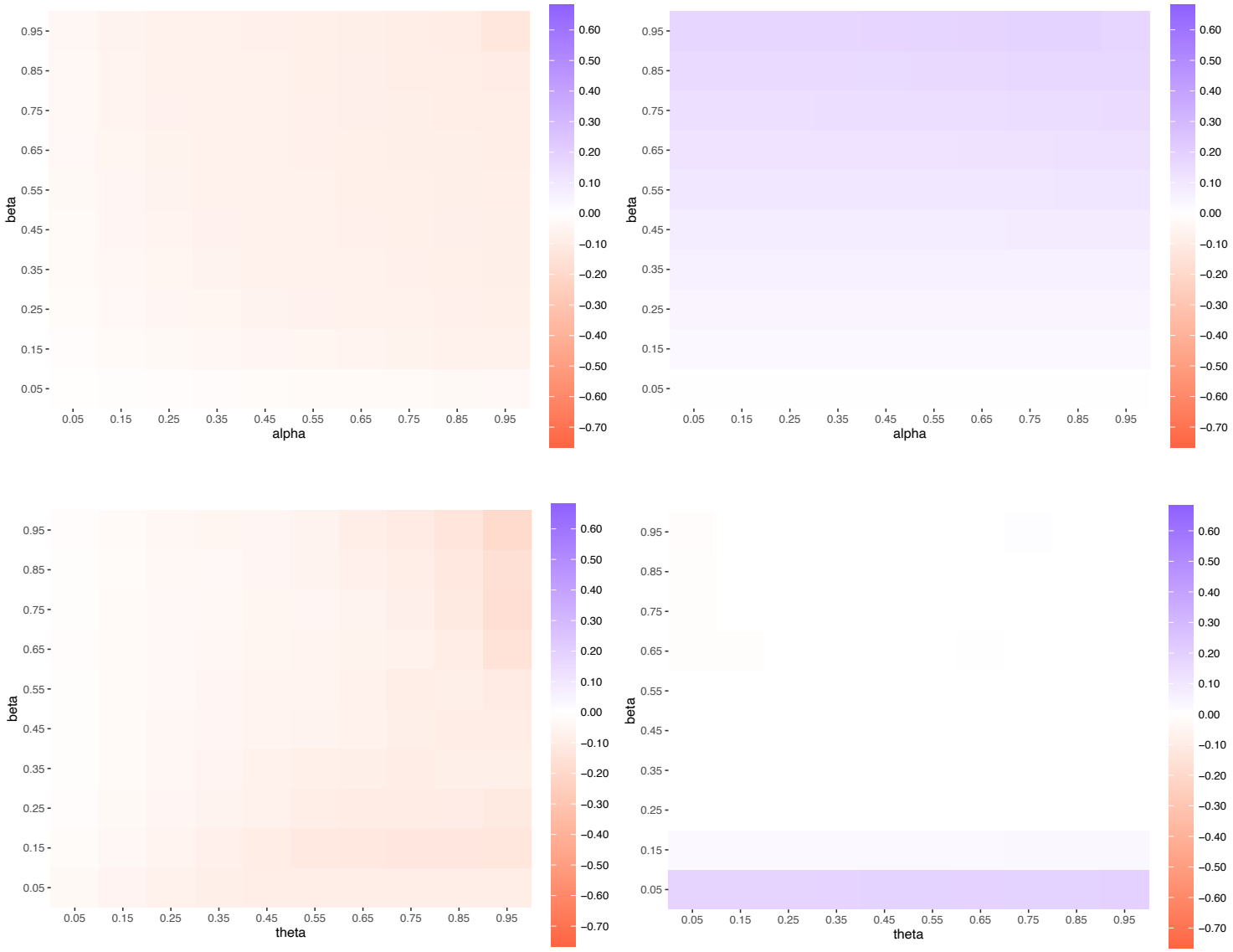
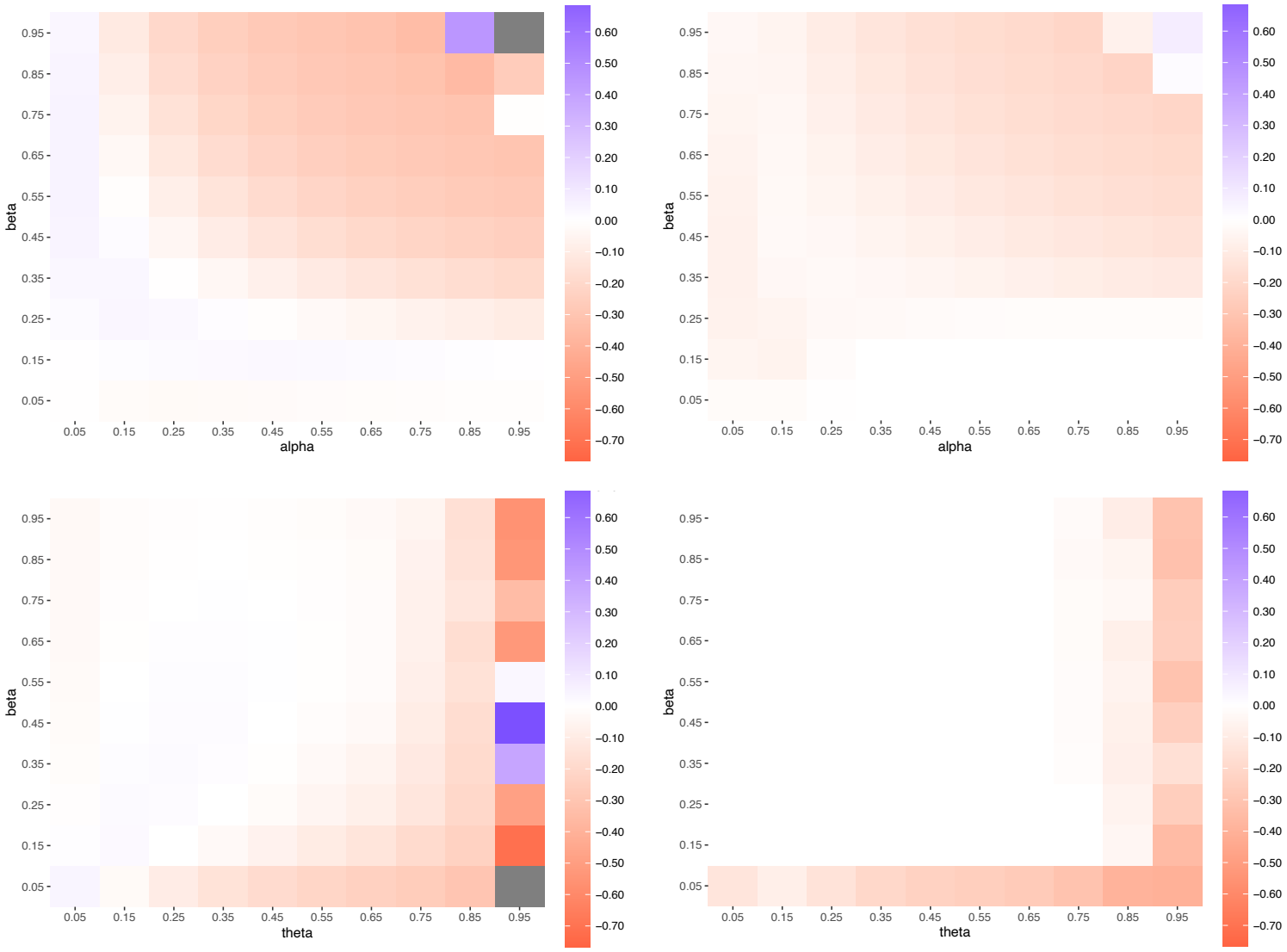

beta
alpha
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
0.60 0.50 0.40 0.30 0.20 0.10 0.00 −0.10 −0.20 −0.30 −0.40 −0.50 −0.60 −0.70
beta
alpha
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
0.60 0.50 0.40 0.30 0.20 0.10 0.00 −0.10 −0.20 −0.30 −0.40 −0.50 −0.60 −0.70
beta
theta
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
0.60 0.50 0.40 0.30 0.20 0.10 0.00 −0.10 −0.20 −0.30 −0.40 −0.50 −0.60 −0.70
beta
theta
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
0.60 0.50 0.40 0.30 0.20 0.10 0.00 −0.10 −0.20 −0.30 −0.40 −0.50 −0.60 −0.70

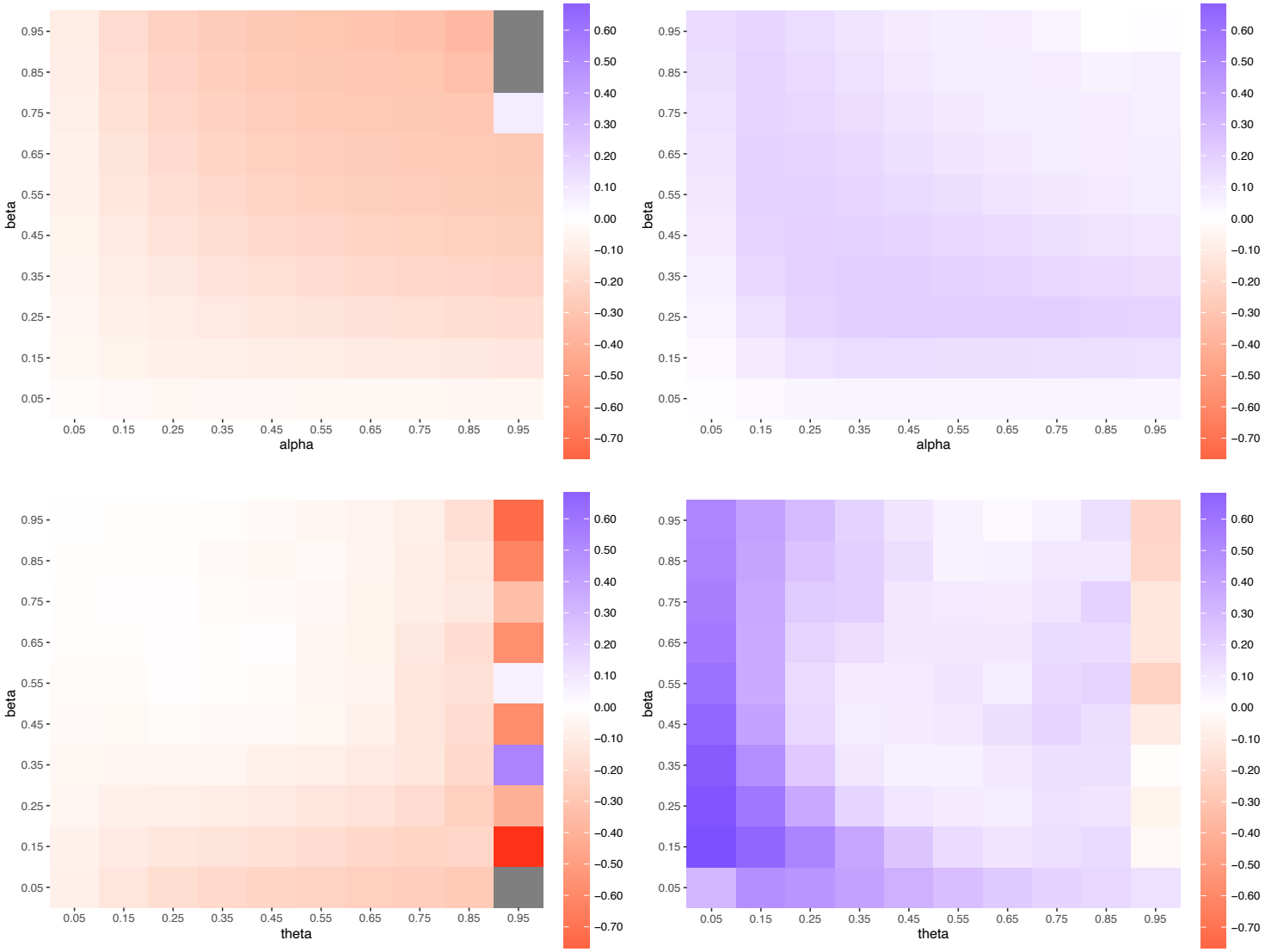
Figure S.20

Differences in associative strength (left panels) and associability (right panels) across different alpha-beta and theta-beta combinations for (a) XY in the 0% group, (b) ZY in the 0% group, (c) XY in the 50% group, and (d) ZY in the 50% group simulated under the summed error learning algorithm with the Mackintosh attention in Experiment 2.3.

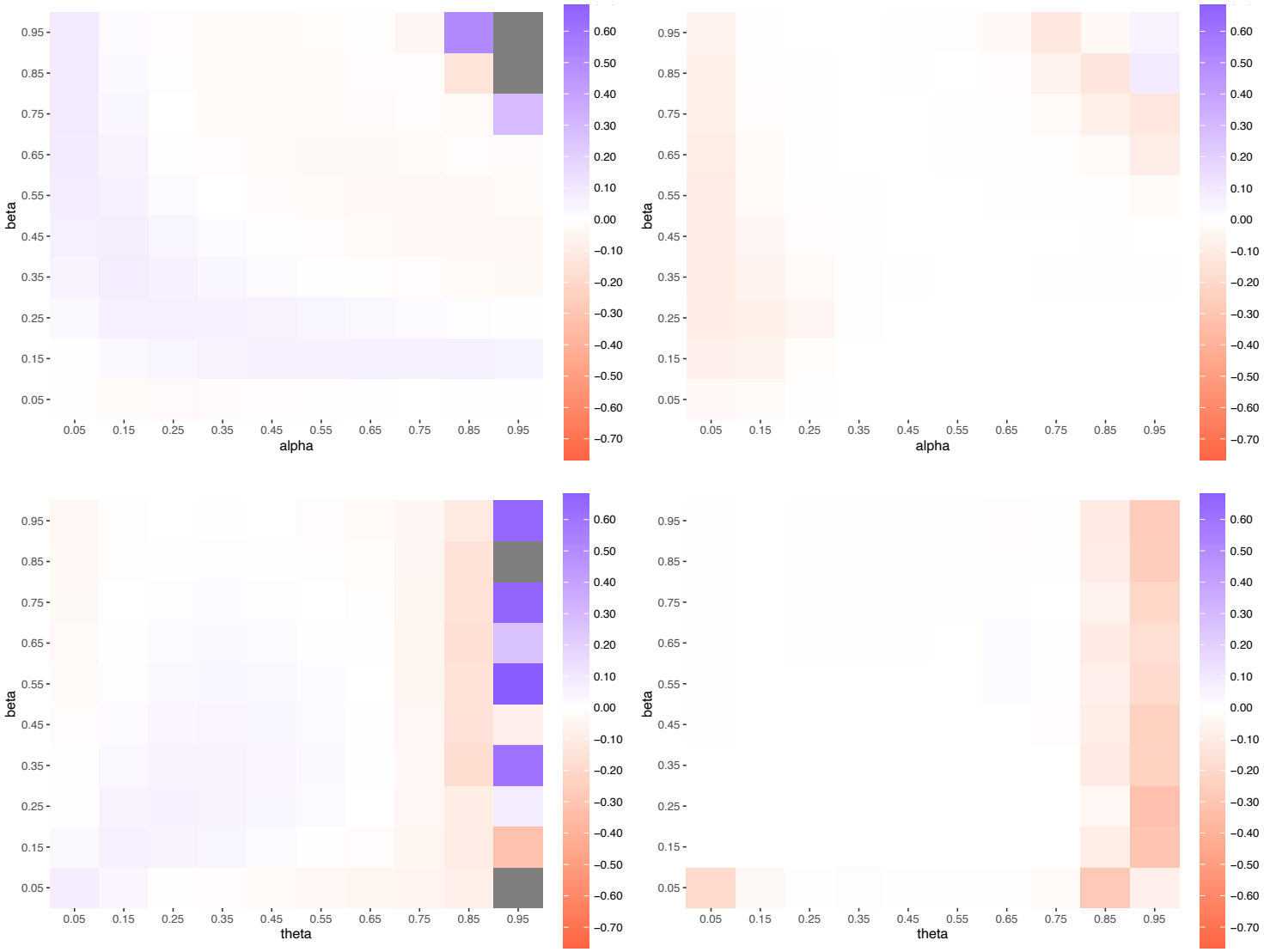
(a)



(d)


beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

(c)


beta
alpha
theta
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

(d)

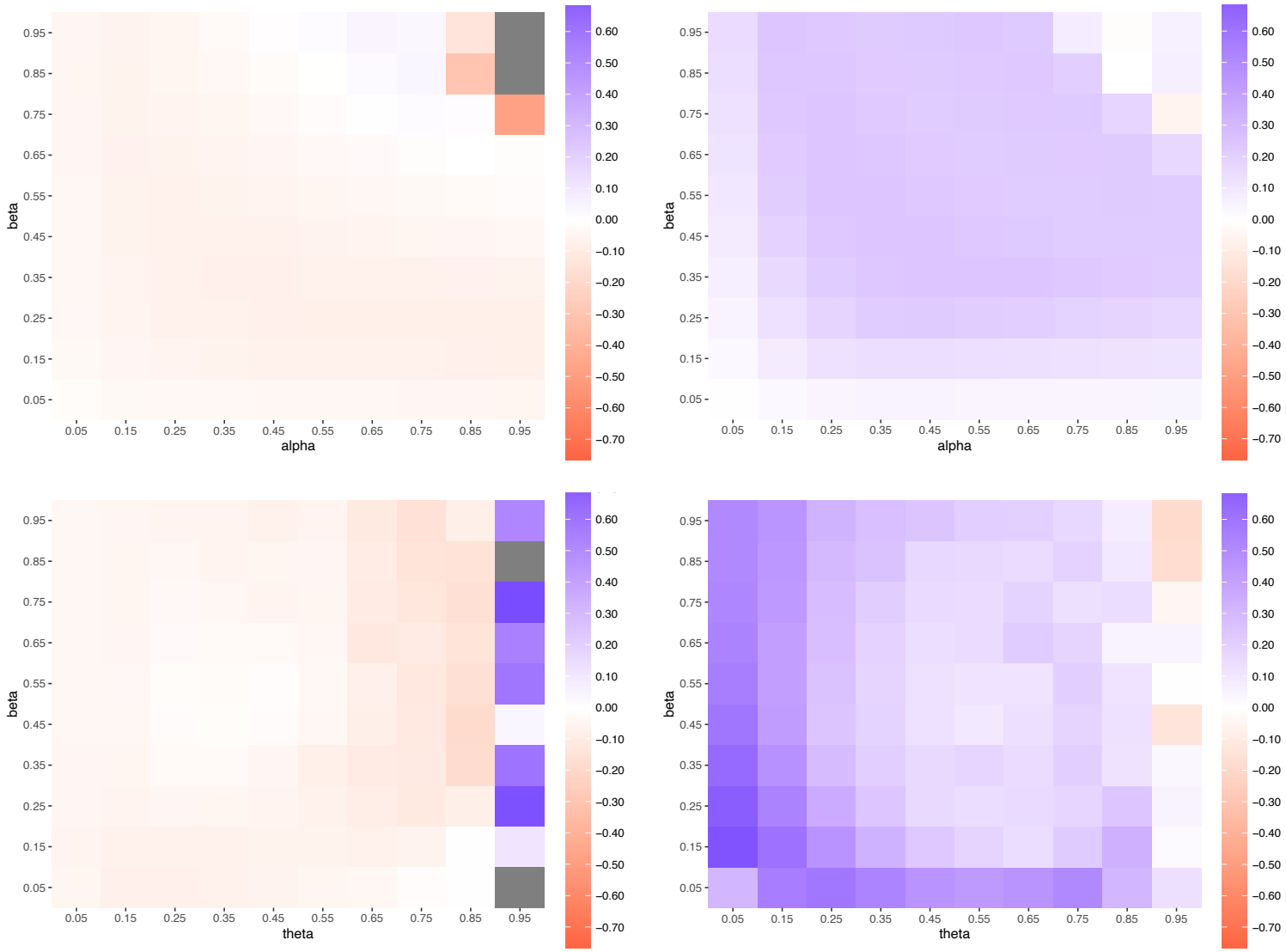
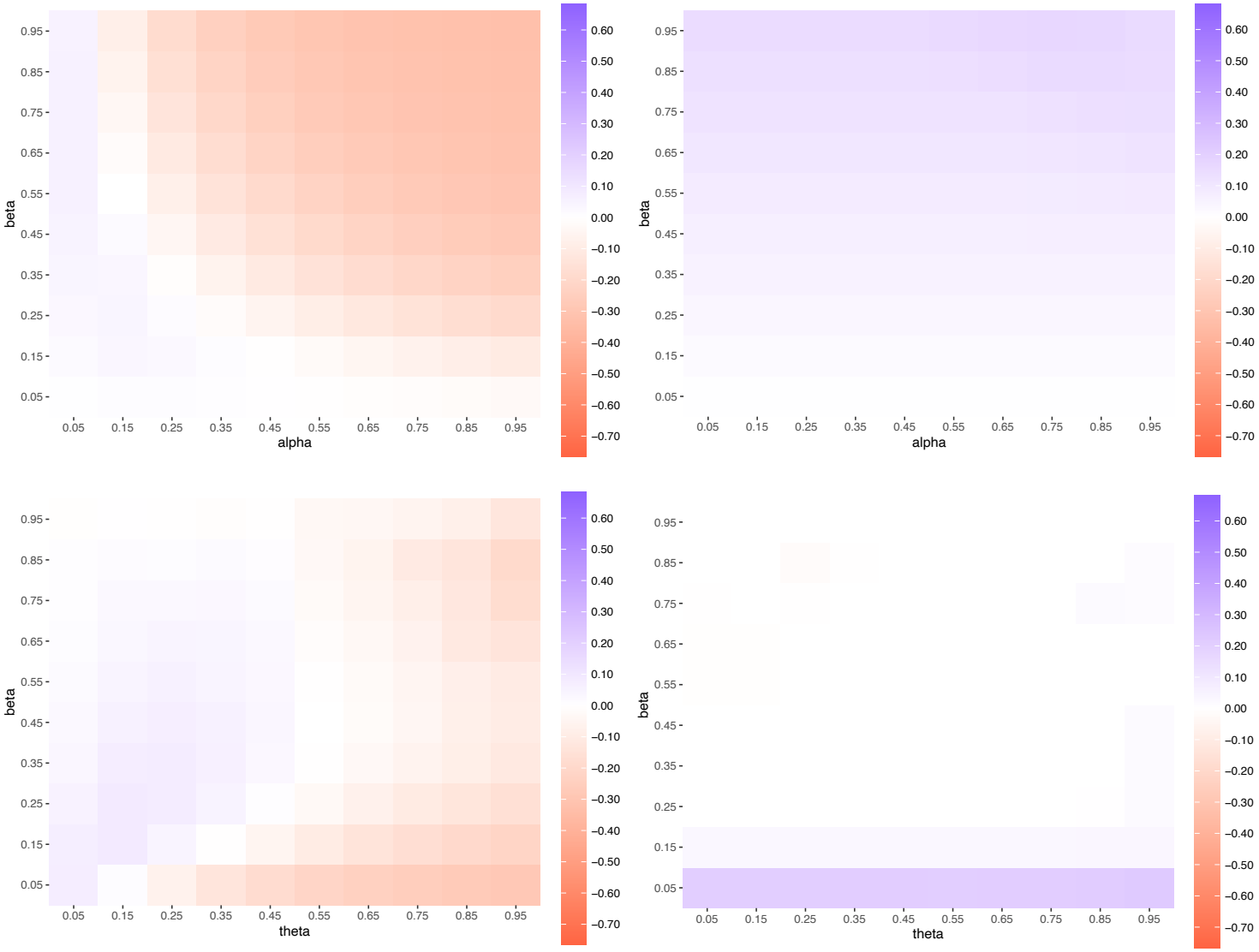

beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

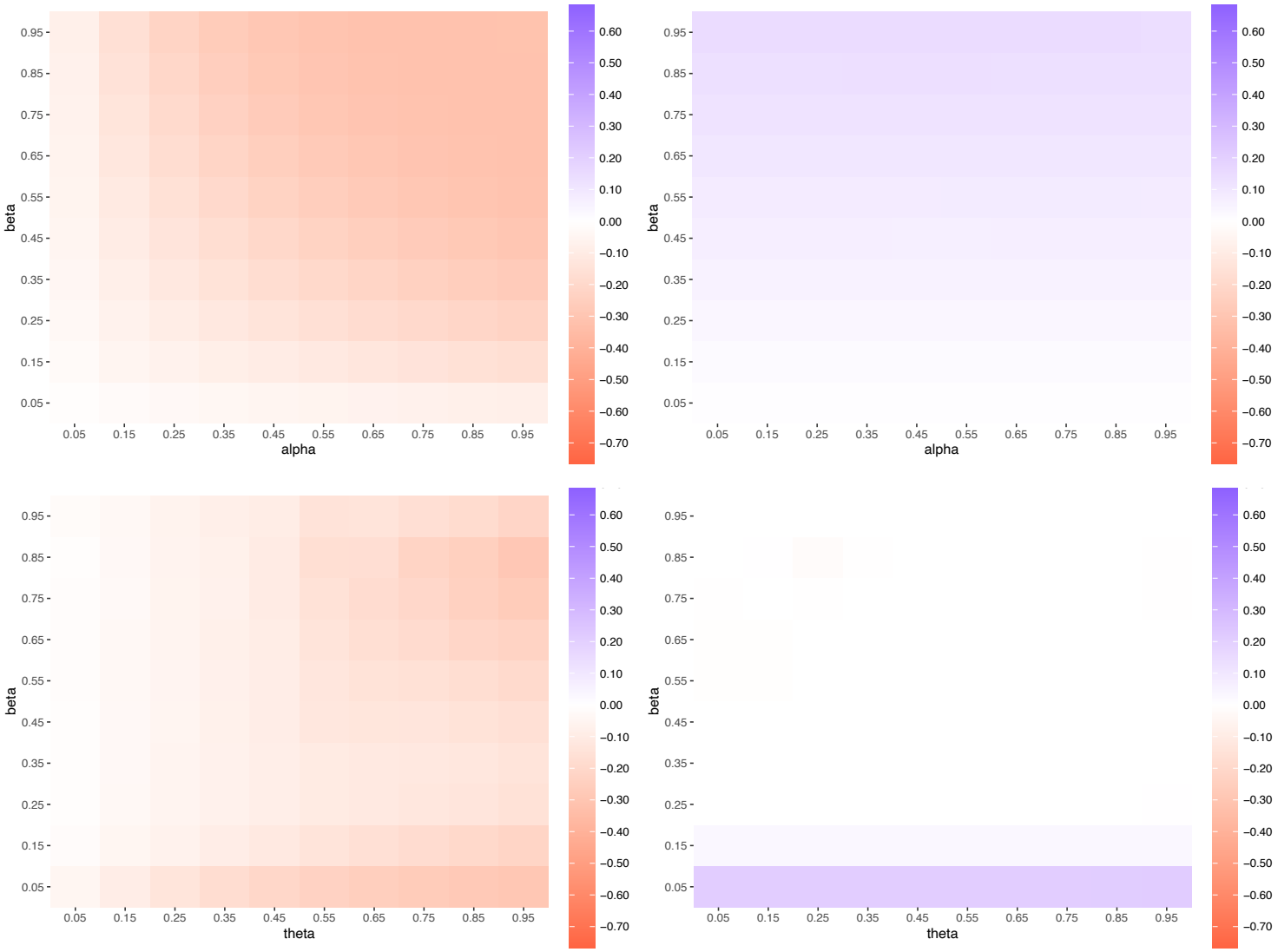
Figure S.21

Differences in associative strength (left panels) and associability (right panels) across different alpha-beta and theta-beta combinations for (a) XY in the 0% group, (b) ZY in the 0% group, (c) XY in the 50% group, and (d) ZY in the 50% group simulated under the summed error learning algorithm with the Uengoer attention in Experiment 2.3.

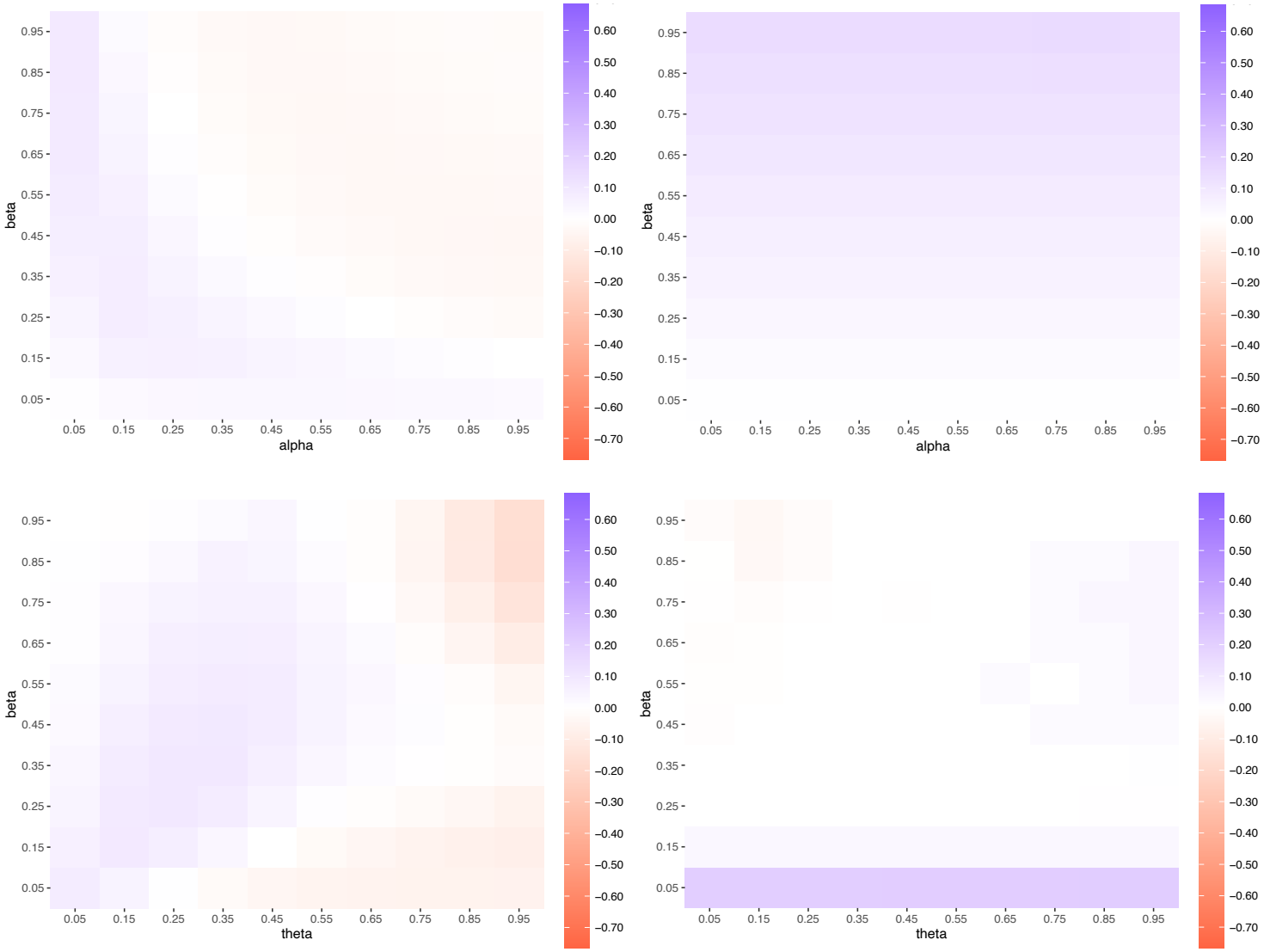
(a)



(b)


beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

(c)


beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

(d)

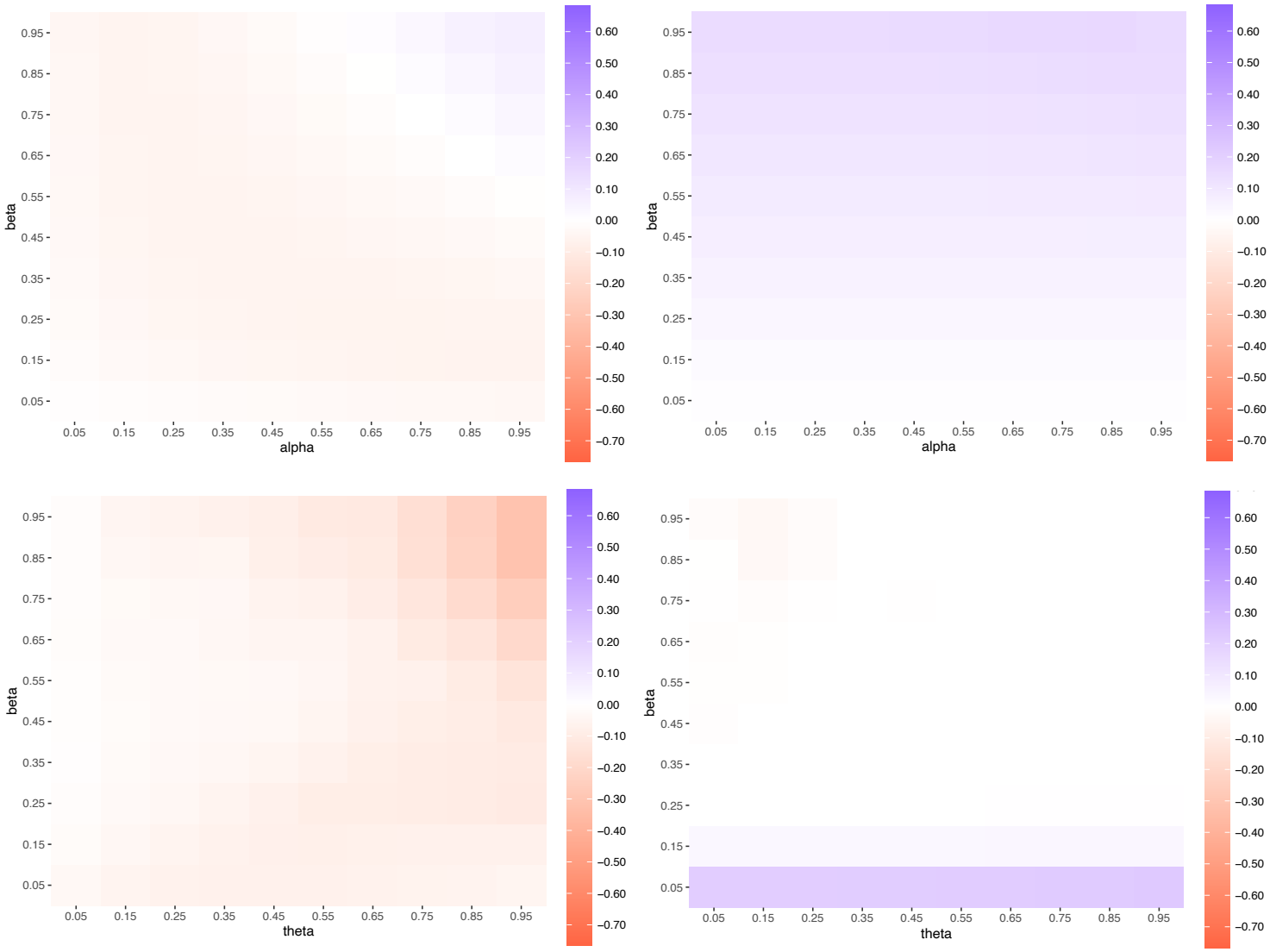
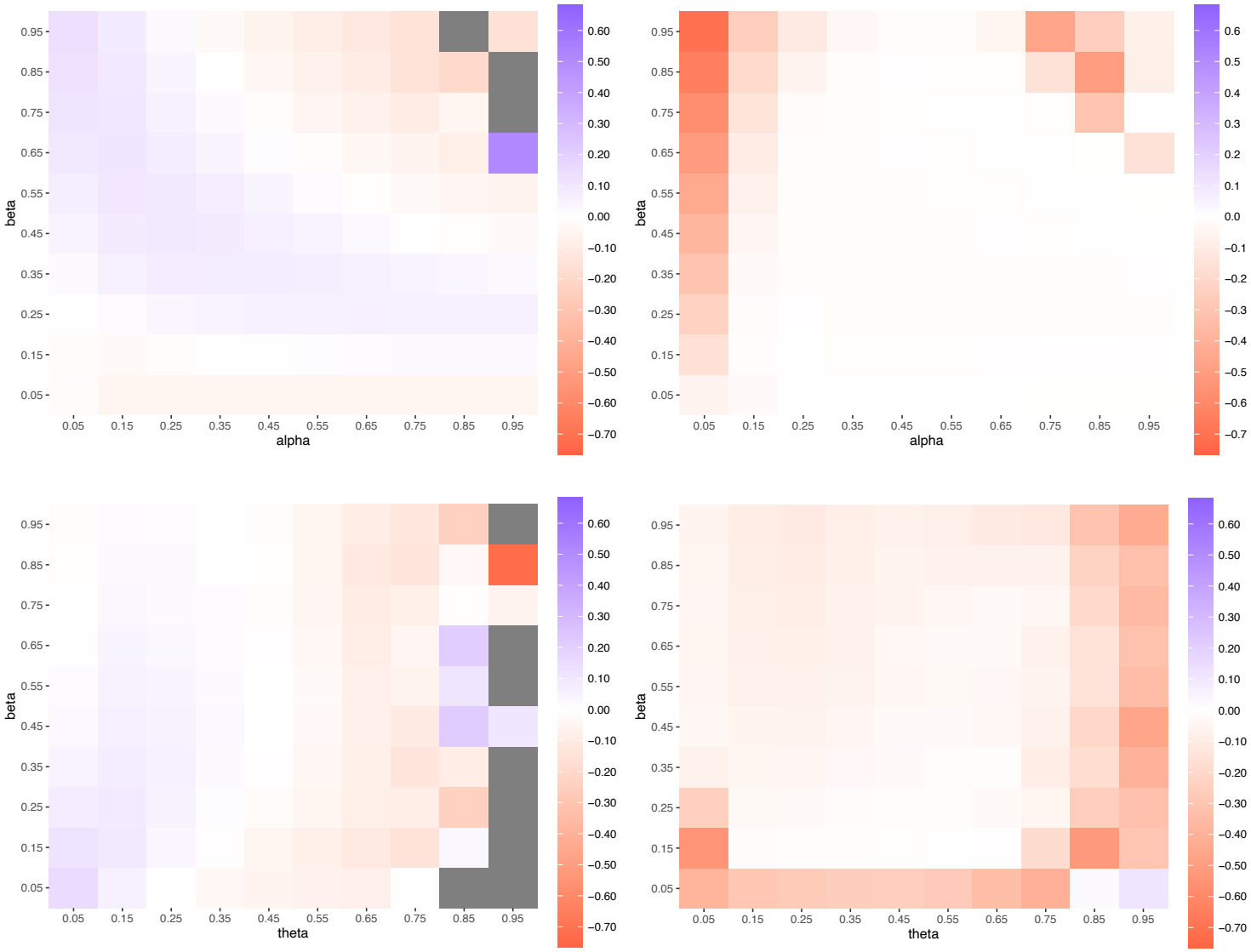

beta
alpha
theta
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

Figure S.22

Differences in associative strength (left panels) and associability (right panels) across different alpha-beta and theta-beta combinations for (a) XY and (b) ZY simulated under the summed error learning algorithm with the Mackintosh attention in Experiment 2.4.

(a)



(b)

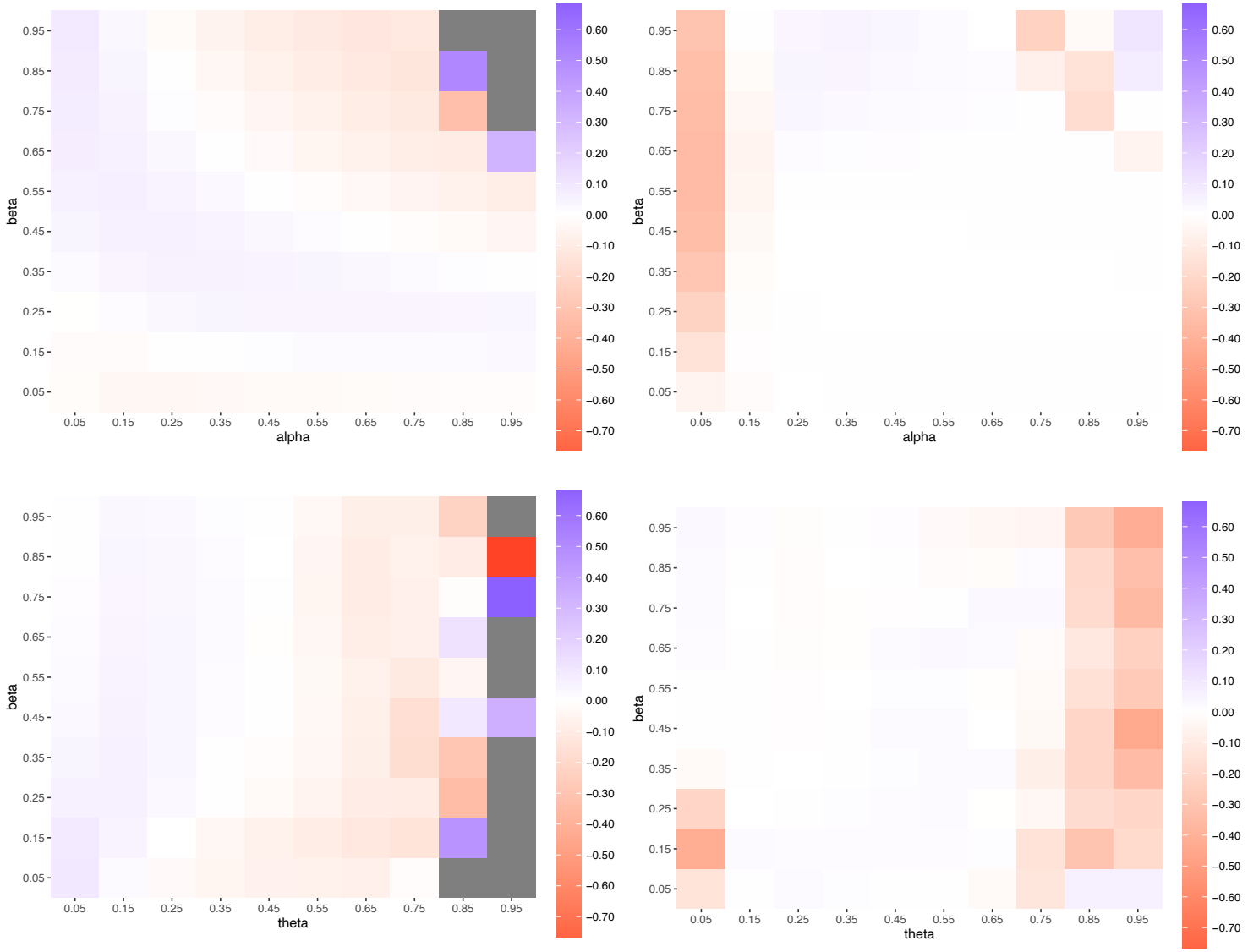
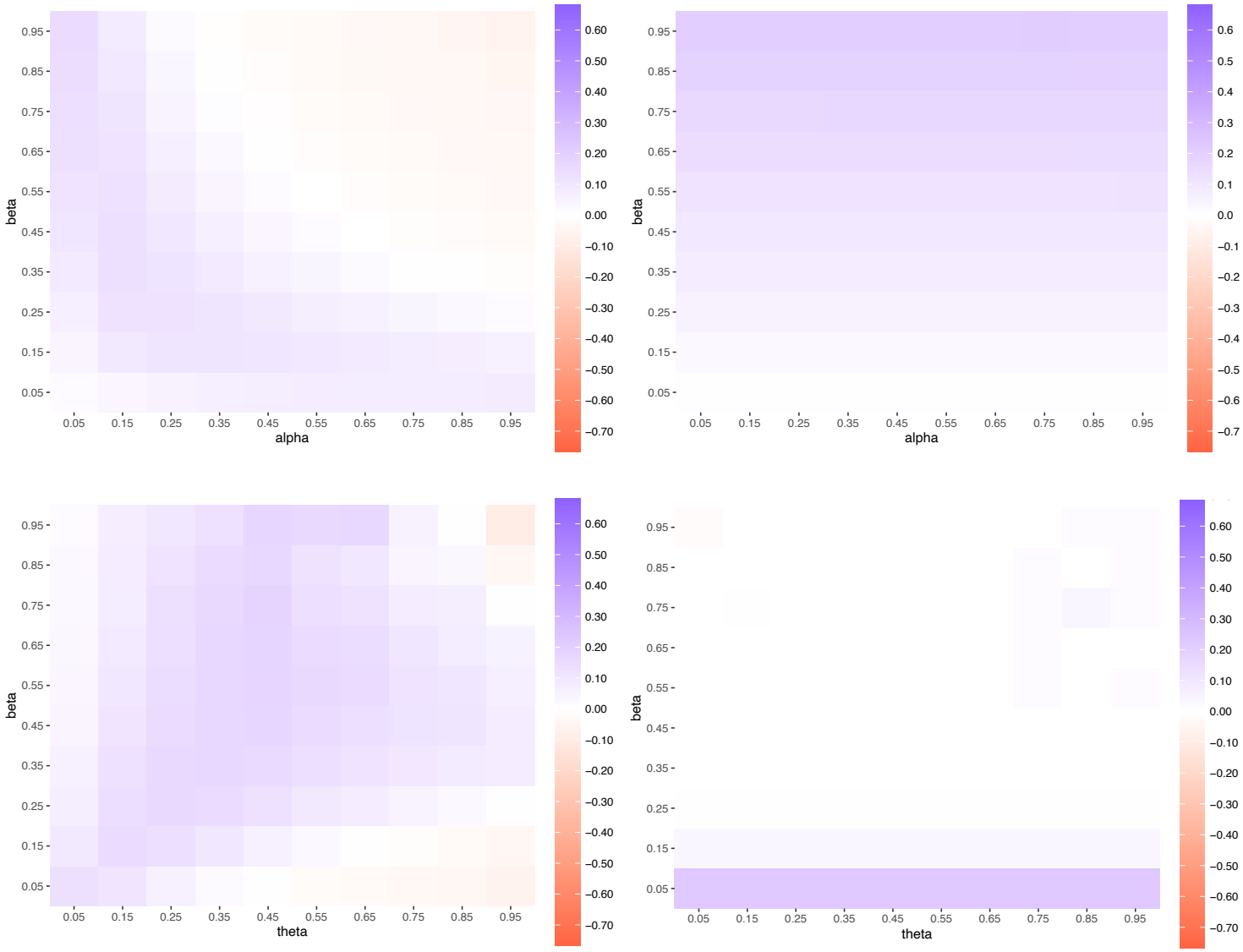

beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

Figure S.23

Differences in associative strength (left panels) and associability (right panels) across different alpha-beta and theta-beta combinations for (a) XY and (b) ZY simulated under the summed error learning algorithm with the Uengoer attention in Experiment 2.4.

(a)



(b)


beta
alpha
theta
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
0.60
0.50
0.40
0.30
0.20
0.10
0.00
−0.10
−0.20
−0.30
−0.40
−0.50
−0.60
−0.70

Supplementary Materials: Chapter 3

Experiment Instructions

Additive Pretraining

If two medicines that produce a mild hormone increase when taken alone are taken together, they WILL produce a larger hormone increase. That is, the medicines will cause a LARGER hormone increase when taken together compared to the increase that each medicine caused when taken separately.

Non-additive Pretraining

If two medicines that produce a mild hormone increase when taken alone are taken together, they will NOT produce a larger hormone increase. That is, the medicines will cause the SAME mild hormone increase when taken together as they each did when taken separately.

Preventative Pretraining

Note that one medicine CAN PREVENT the effect of another medicine. That is, if one medicine causes a hormone increase, another medicine can potentially prevent this increase from happening.

Non-Preventative Pretraining

Note that one medicine CANNOT PREVENT the effect of another medicine. That is, if one medicine causes a hormone increase, another medicine cannot prevent this increase from happening.

Forced Choice Test

Finally, we are going to ask you to choose among pairs of different medicines based on what you have learned so far. For each pair, you will first be asked to choose which medicine you think is more likely to result in a hormone increase in Patient X. Then you will be asked to indicate which medicine you feel more confident about judging. In other words, you will be asked to choose which medicine you are most sure either does or does not cause a hormone increase. For this second question, it does not matter whether you think each medicine causes a hormone increase and does not cause a hormone increase. You merely need to choose the one that you feel most confident judging.

Training Curves

Experiment 3.1

Figure S.24

Mean proportion of hormone increase predictions for all ten trial types across eight blocks of training in Experiment 3.1. Error bars indicate standard error of mean (SEM).

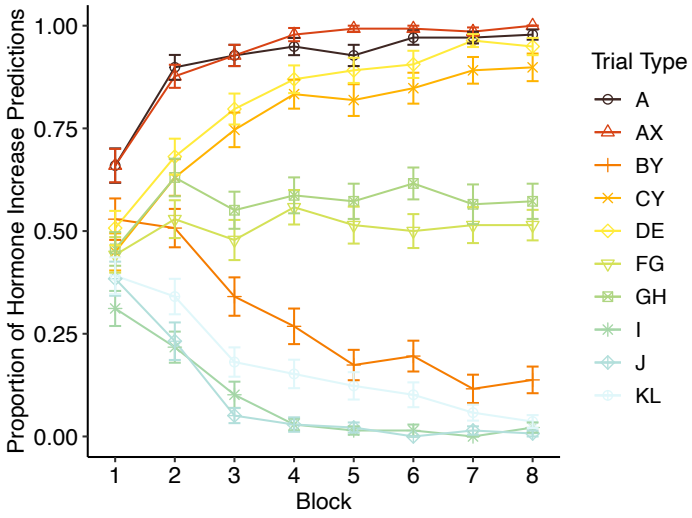
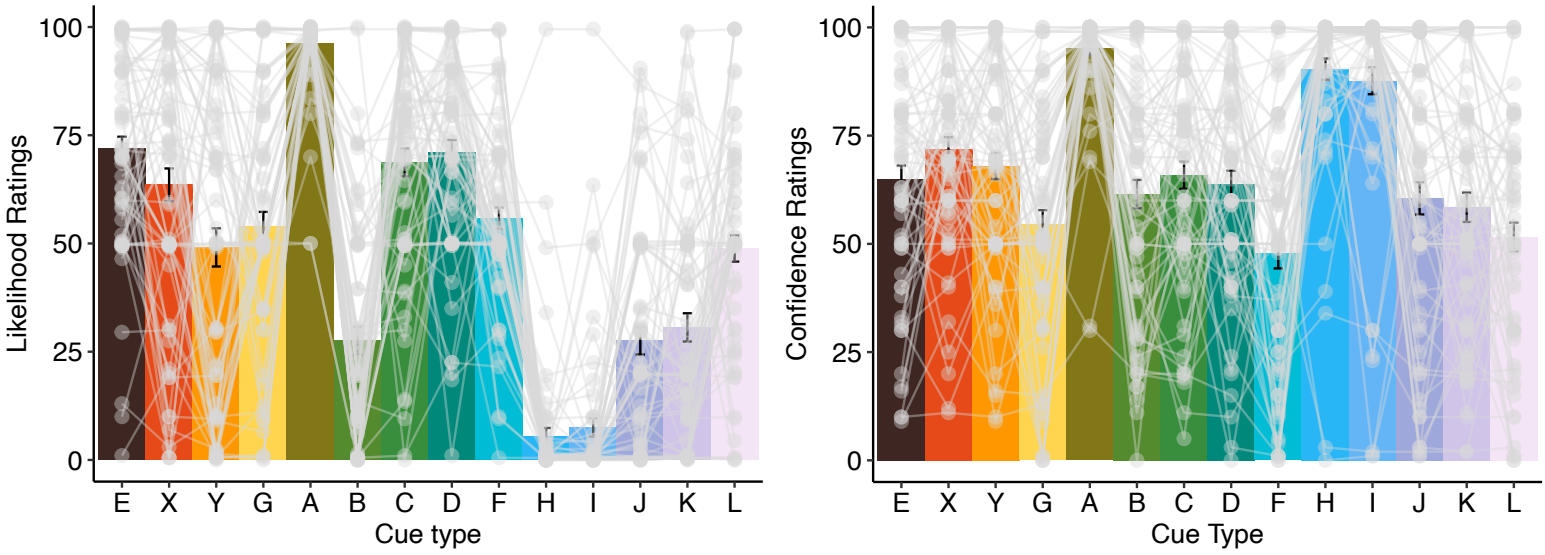


Figure S.25

(a) Mean likelihood ratings and (b) Mean confidence ratings for all fourteen cue types in Experiment 3.1. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participants.

(a) (b)



Experiment 3.2

Figure S.26

Mean proportion of hormone increase predictions for all six trial types across four blocks of pretraining for (a) the additive group and (b) the non-additive group in Experiment 3.2. Error bars indicate standard error of mean (SEM).

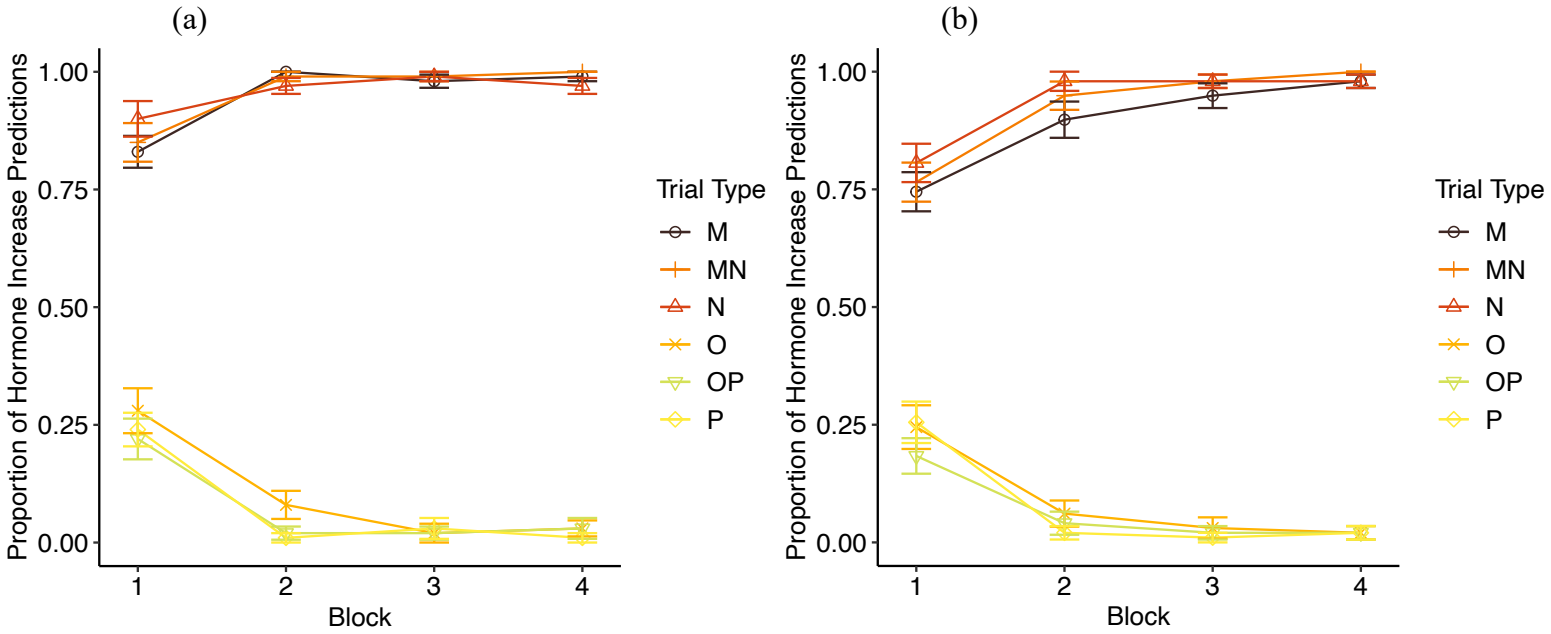


Figure 3

Mean proportion of hormone increase predictions for all ten trial types across eight blocks of main training for (a) the additive group and (b) the non-additive group in Experiment 3.2. Error bars indicate standard error of mean (SEM).

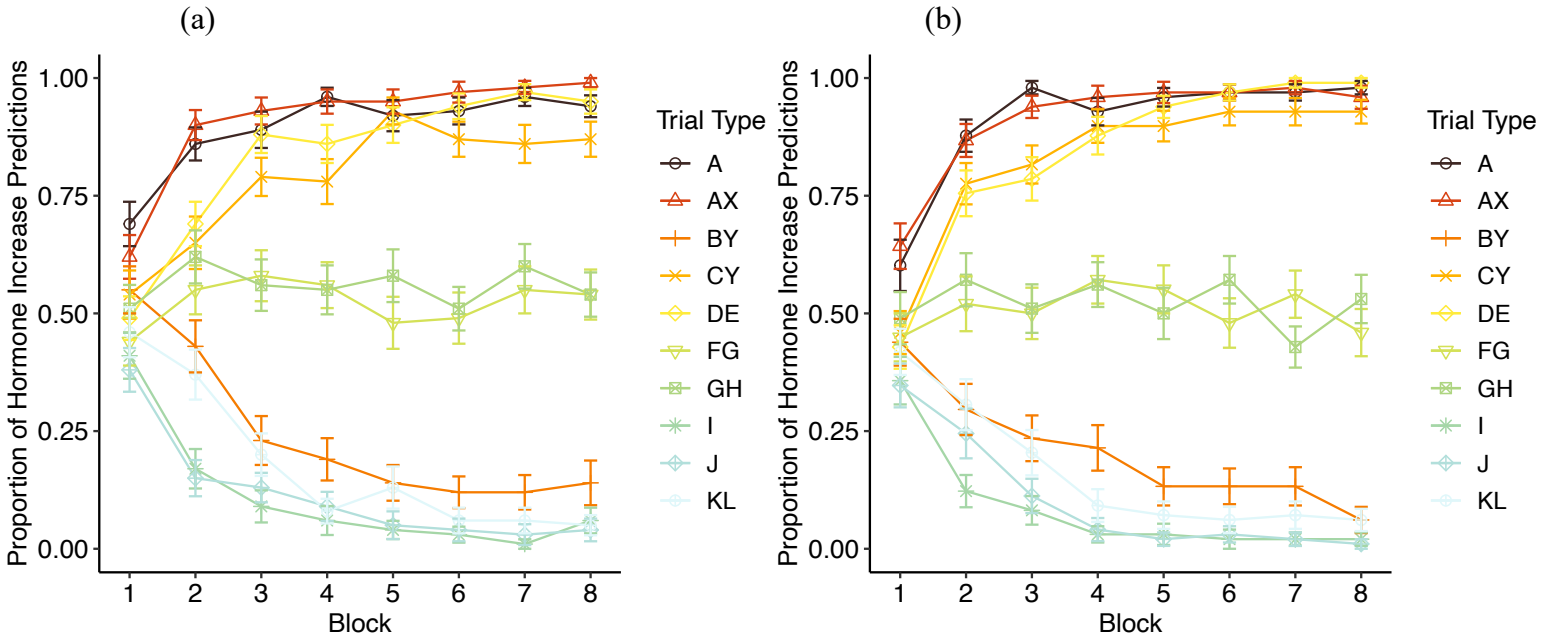
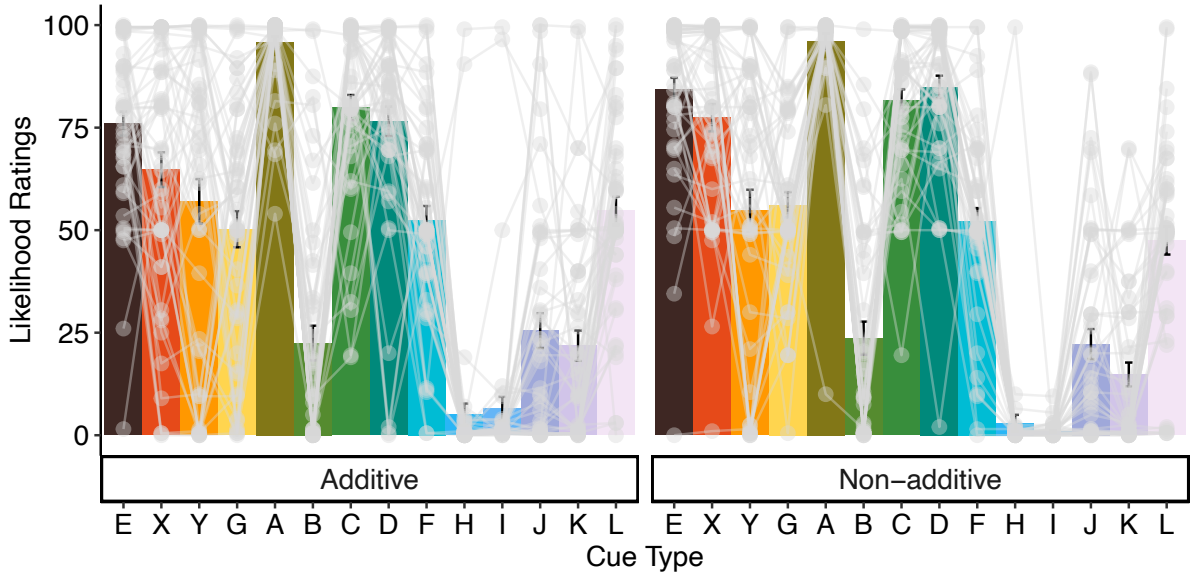


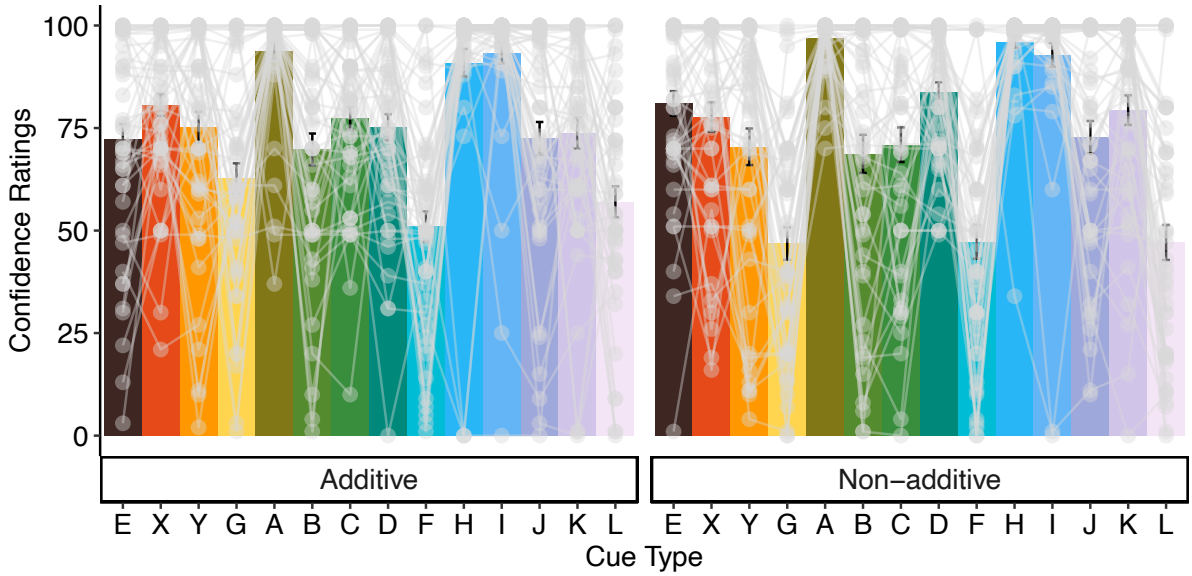
Figure S.27

(a) Mean likelihood ratings and (b) Mean confidence ratings on the ratings test all for fourteen cue types in the additive group and the non-additive group of Experiment 3.2. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participants.

(a)



(b)



Experiment 3.3

Figure S.28

Mean proportion of hormone increase predictions for all six trial types across four blocks of pretraining for (a) the non-preventative group and (b) the preventative group in Experiment 3.3. Error bars indicate standard error of mean (SEM).

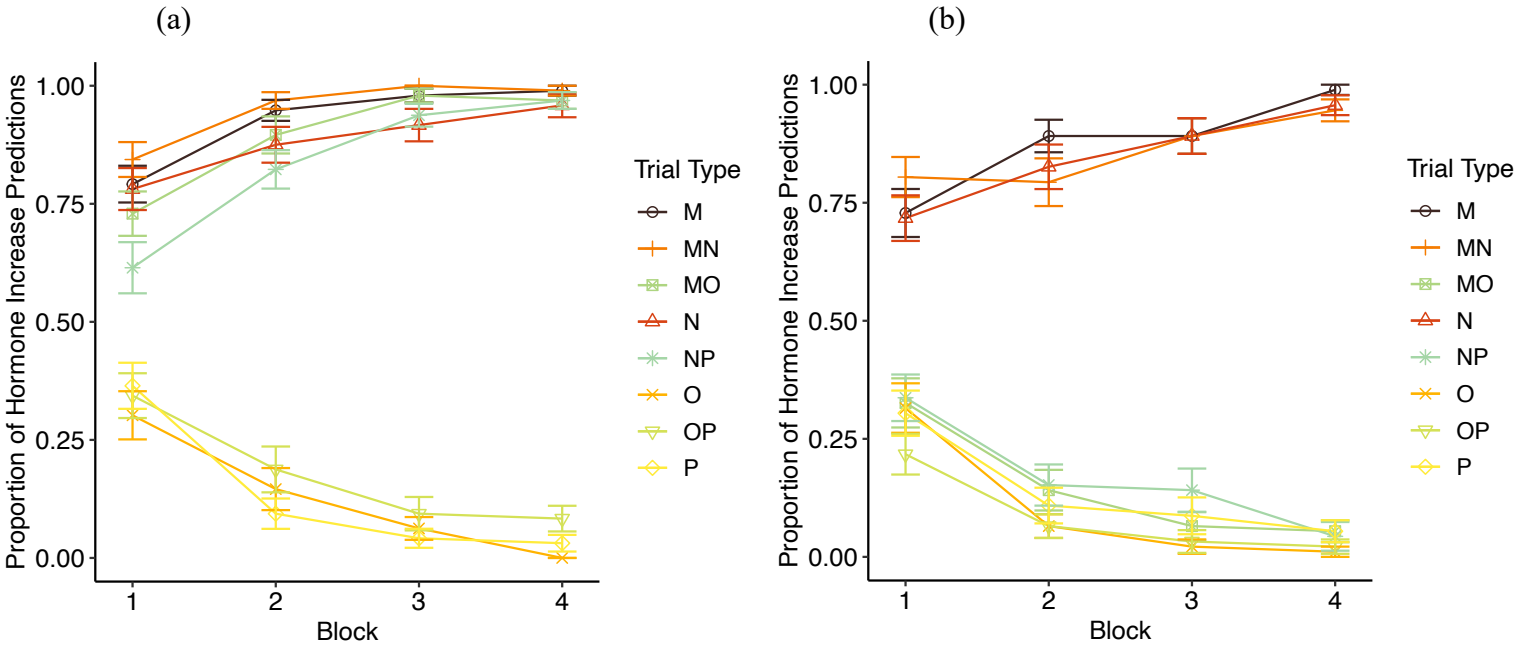


Figure S.29

Mean proportion of hormone increase predictions for all ten trial types across eight blocks of main training for (a) the non-preventative group and (b) the preventative group in Experiment 3.3. Error bars indicate standard error of mean (SEM).

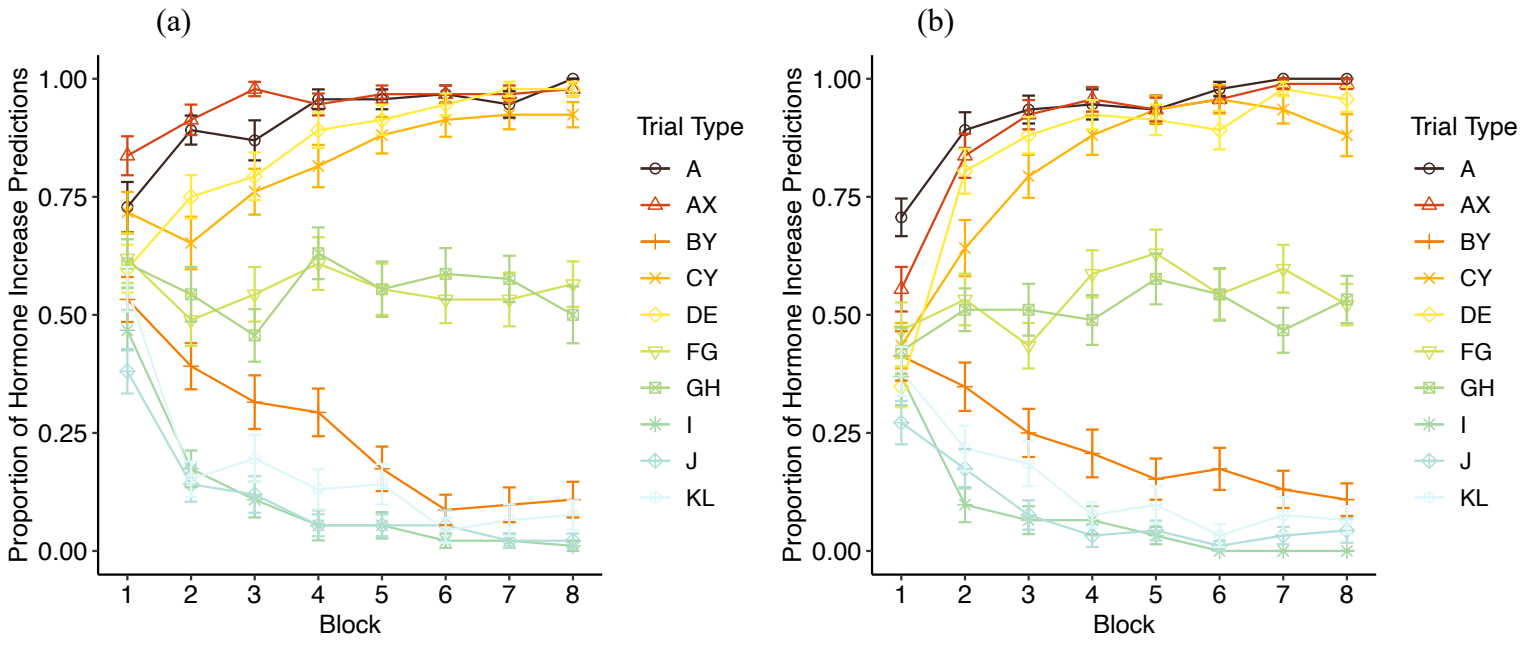
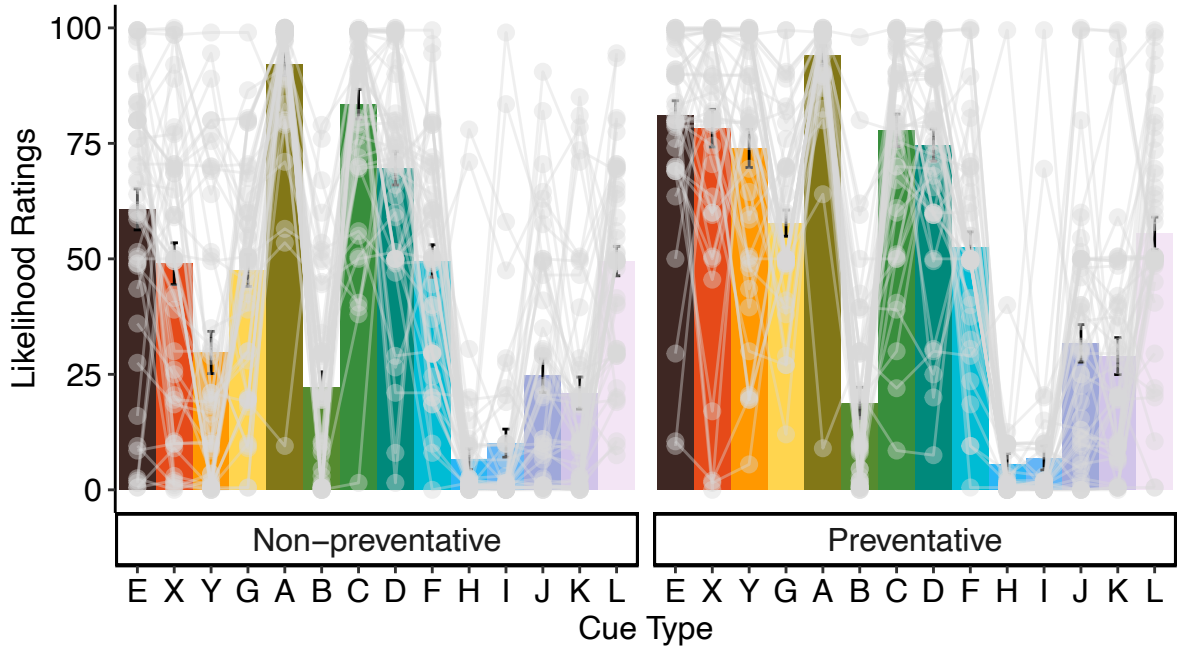
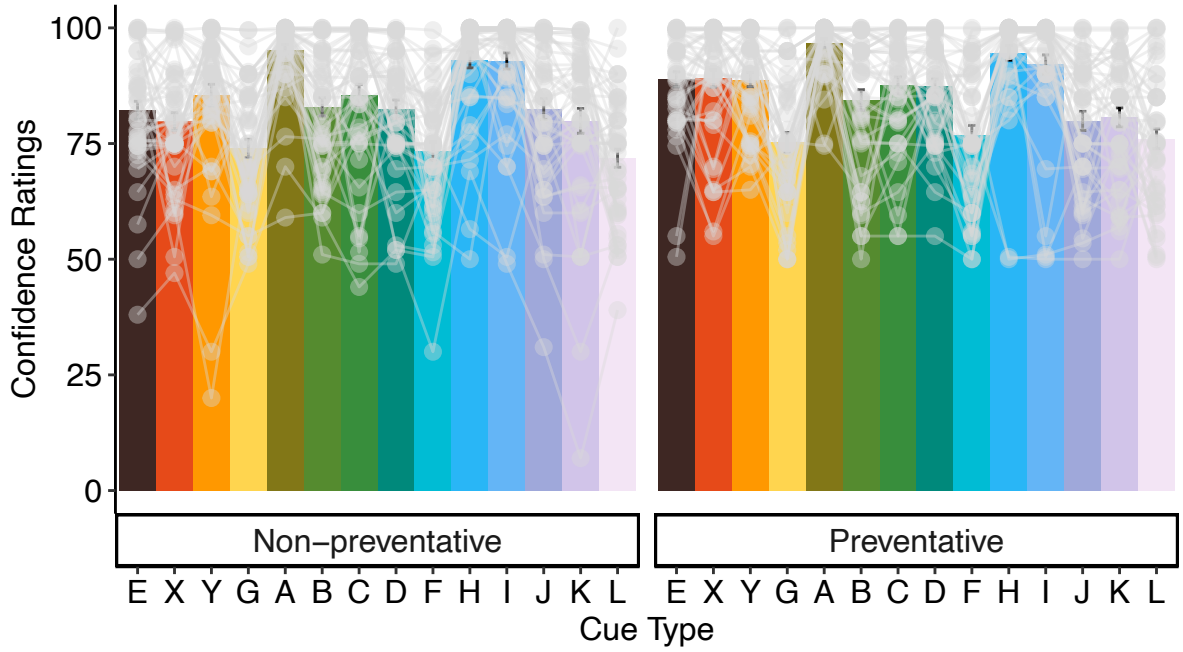


Figure S.30

(a) Mean likelihood ratings and (b) Mean confidence ratings on the ratings test for all fourteen cue types in the non-preventative group and the preventative group of Experiment 3.3. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participants.

(a)





Supplementary Materials: Chapter 4

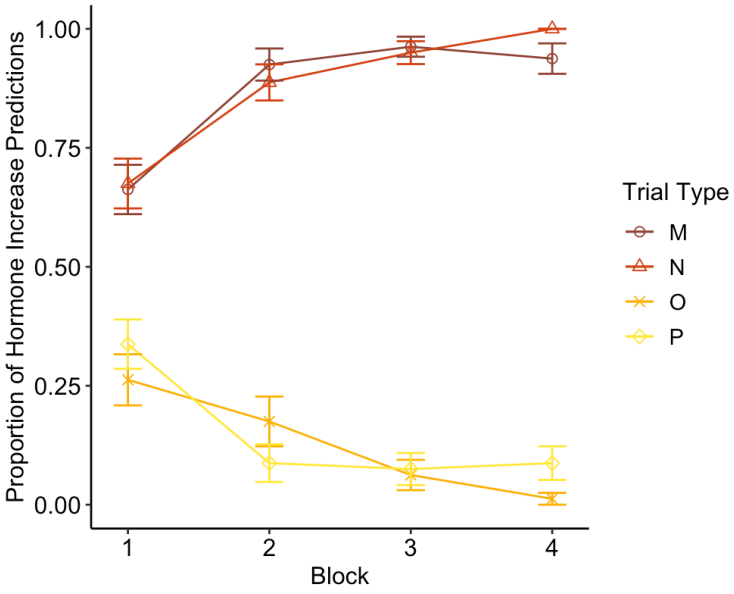
Training Curves

Experiment 4.1

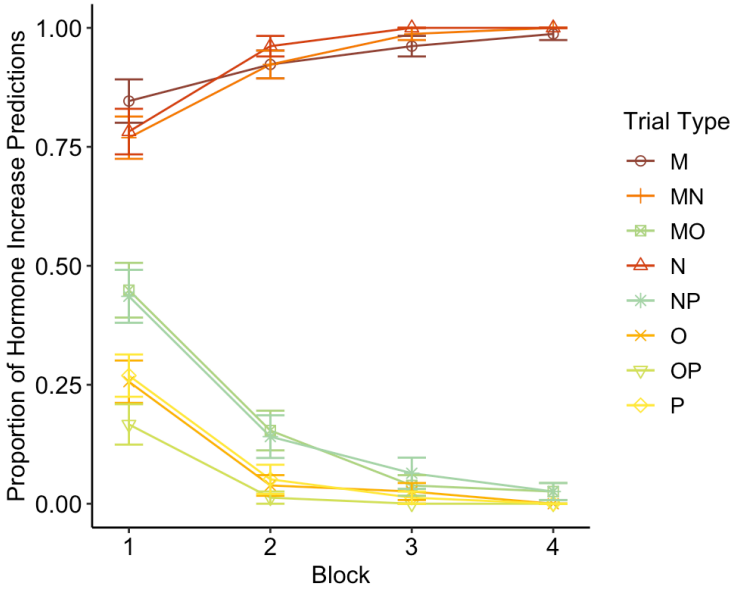
Figure S.31

Proportion of hormone increase predictions across blocks of pretraining in Experiment 4.1 for (a) the neutral group, (b) the additive and preventative group, and (c) the non-additive and non-preventative group. Error bars indicate standard error of mean (SEM).

(a)



(b)



(c)

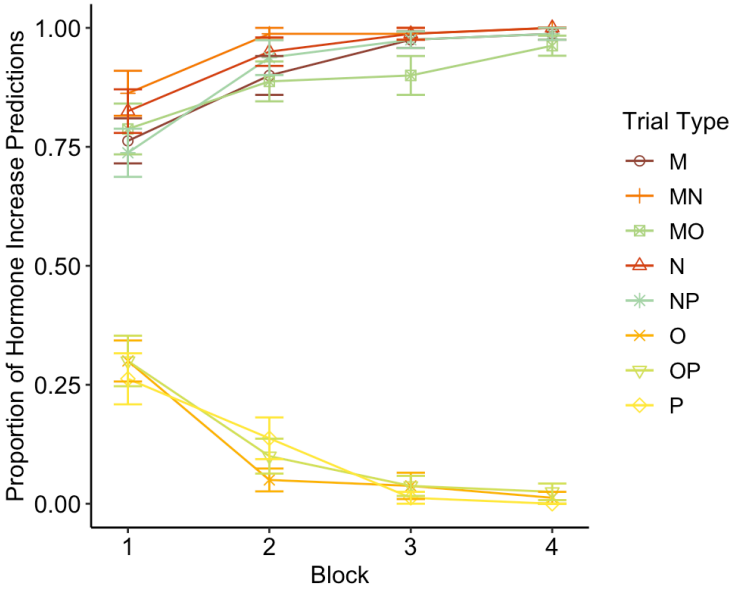


Figure S.32

Proportion of hormone increase predictions across blocks of Stage 1 training in Experiment 4.1 for (a) the neutral group, (b) the additive and preventative group, and (c) the non-additive and non-preventative group. Error bars indicate standard error of mean (SEM).

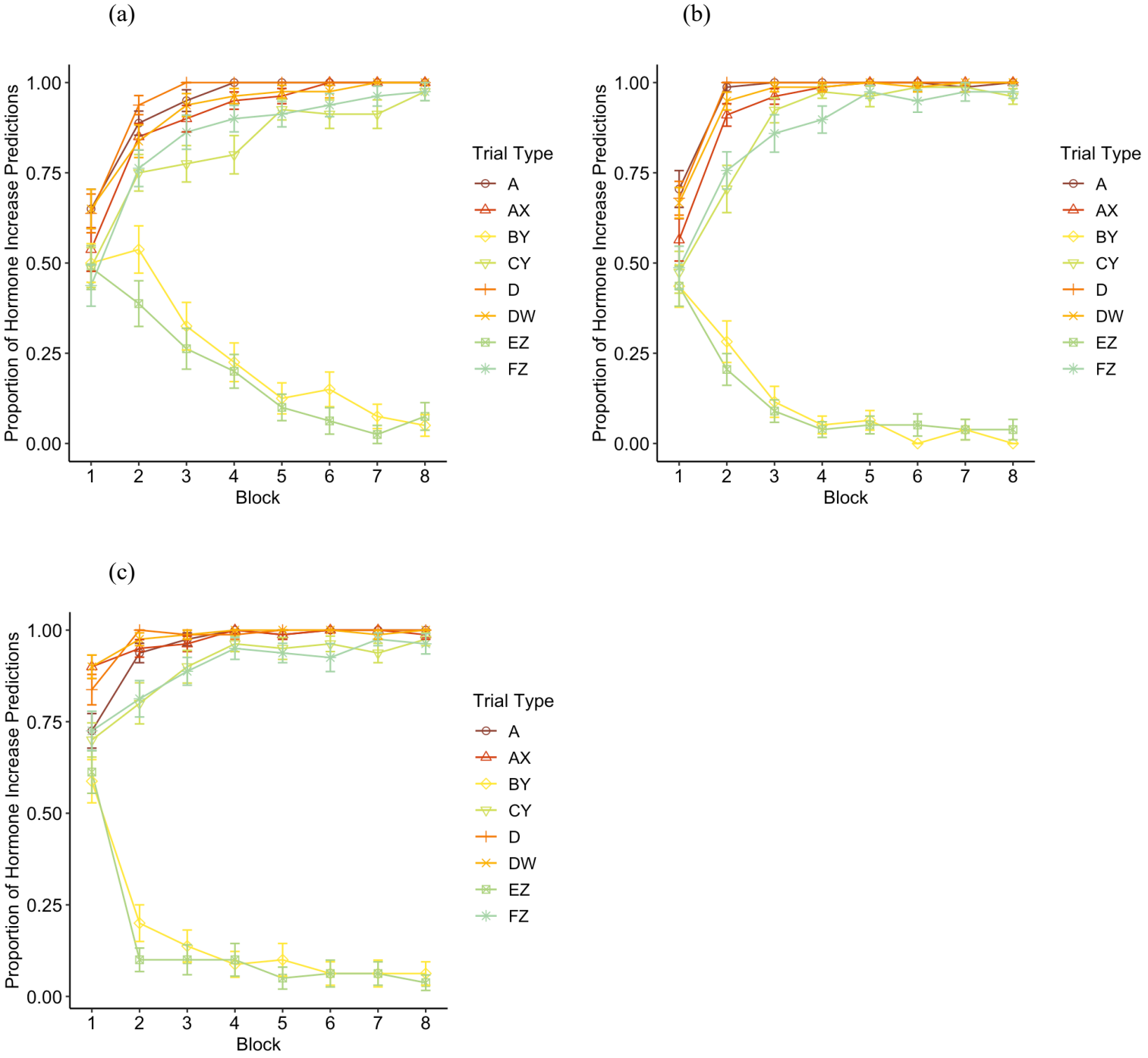
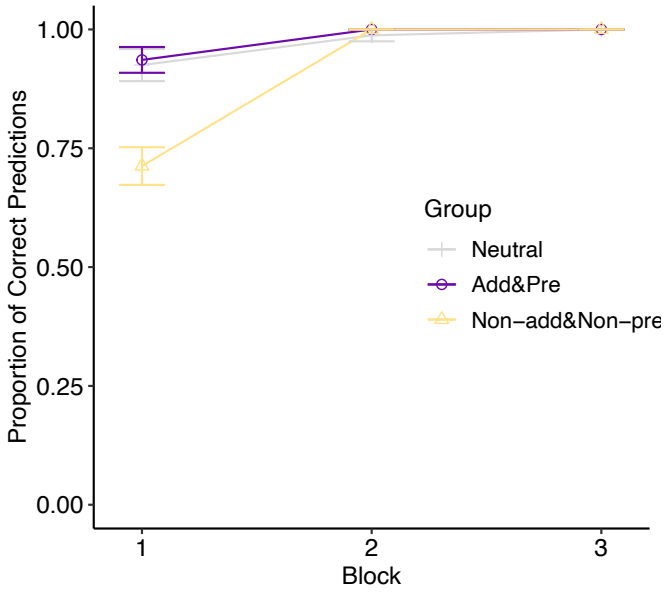


Figure S.33

Proportion of hormone increase predictions across blocks of Stage 2 compound training in Experiment 4.1 for the neutral group, the additive and preventative group, and the non-additive and non-preventative group. Error bars indicate standard error of mean (SEM).



Experiment 4.2

Figure S.34

Proportion of hormone increase predictions across blocks of pretraining in Experiment 4.2 for (a) the additive and preventative group and (b) the non-additive and non-preventative group. Error bars indicate standard error of mean (SEM).

(a) (b)

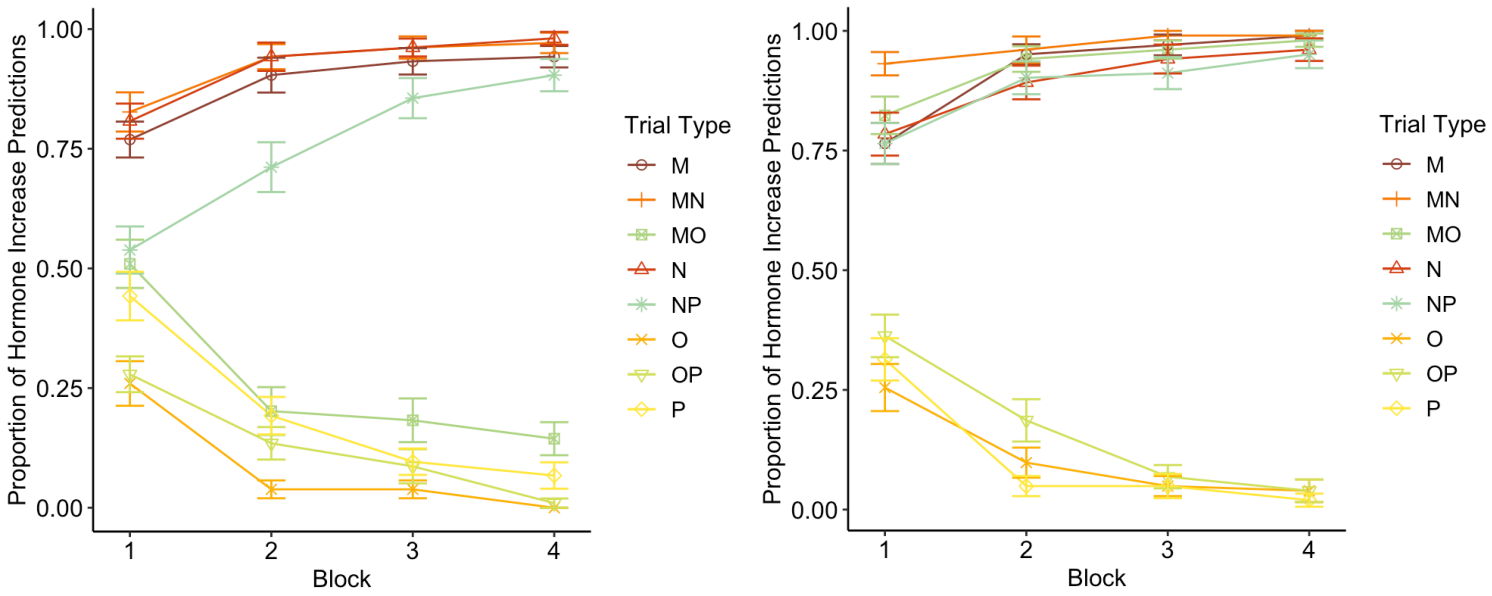


Figure S.35

Proportion of hormone increase predictions across blocks of Stage 1 main training in Experiment 4.2 for (a) the additive and preventative group and (b) the non-additive and non-preventative group. Error bars indicate standard error of mean (SEM).

(a) (b)

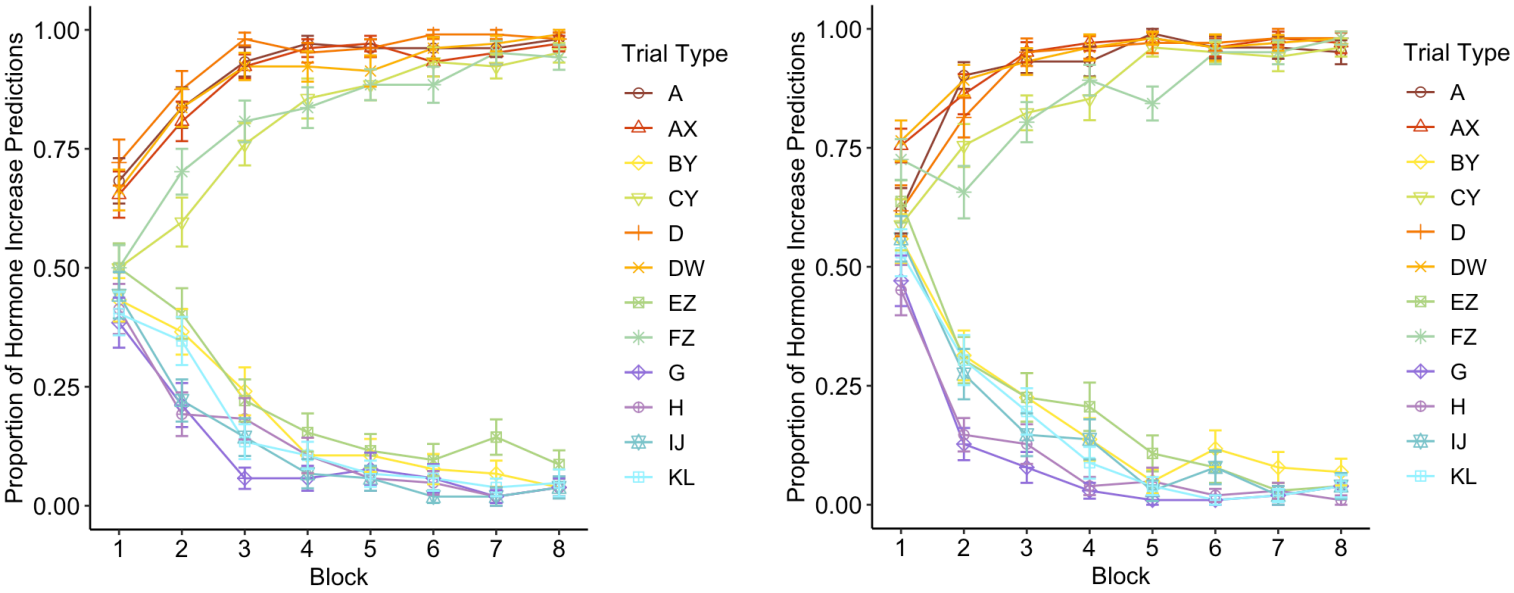
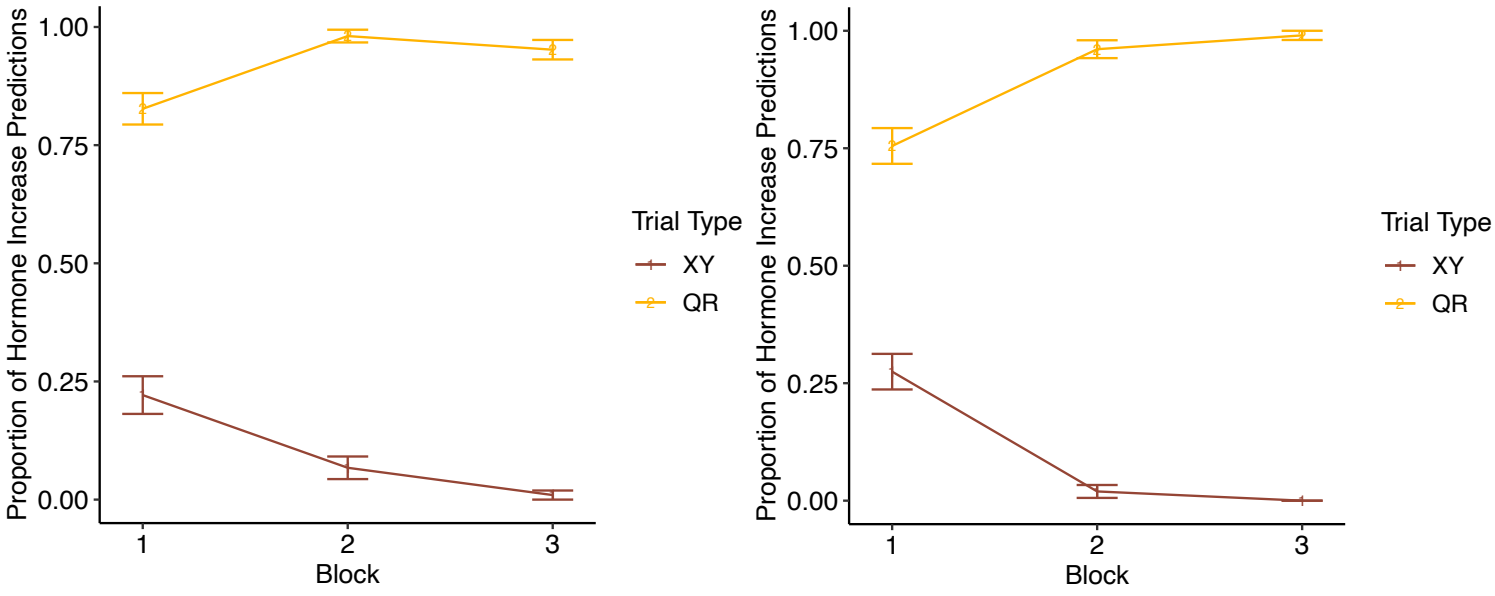


Figure S.36

Proportion of hormone increase predictions across blocks of Stage 2 compound training in Experiment 4.2 for (a) the additive and preventative group and (b) the non-additive and non-preventative group. Error bars indicate standard error of mean (SEM).

(a) (b)



Experiment 4.3

Figure S.37

Proportion of hormone increase predictions across blocks of pretraining in Experiment 4.3 for (a) the additive group and (b) the non-additive group. Error bars indicate standard error of mean (SEM).

(a) (b)

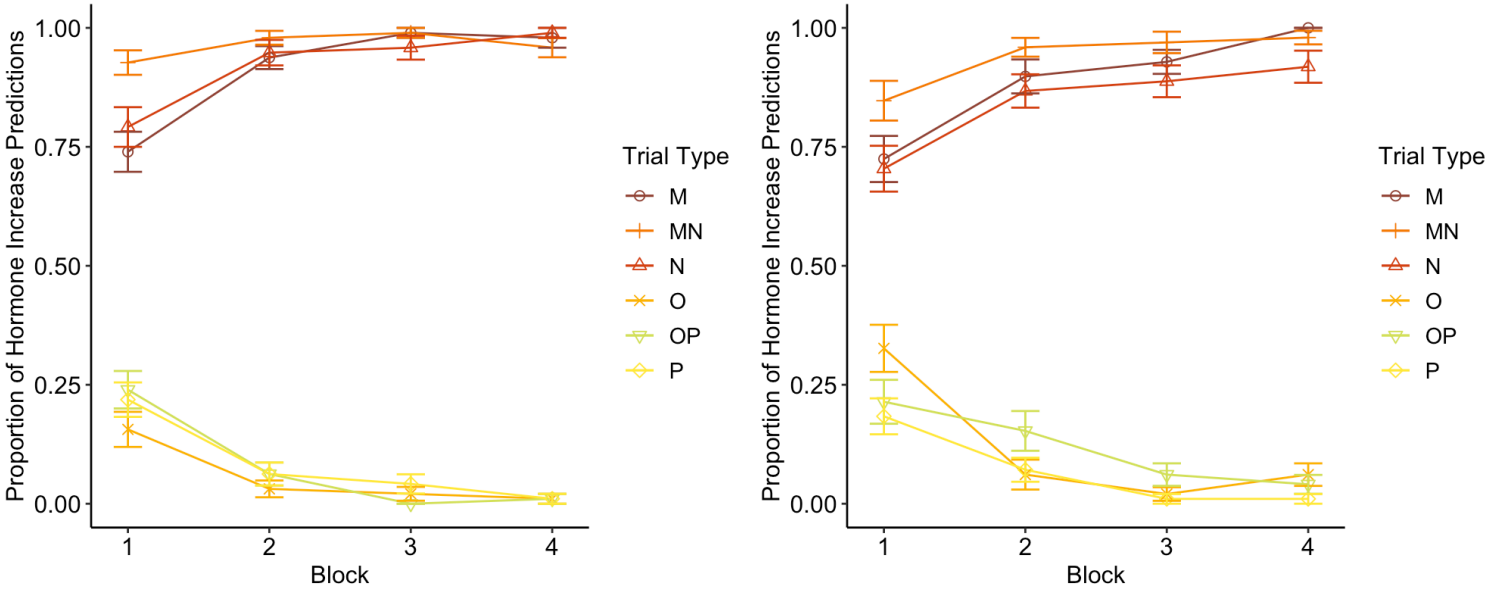


Figure S.38

Proportion of hormone increase predictions across blocks of Stage 1 training in Experiment 4.3 for (a) the additive group and (b) the non-additive group. Error bars indicate standard error of mean (SEM).

(a) (b)

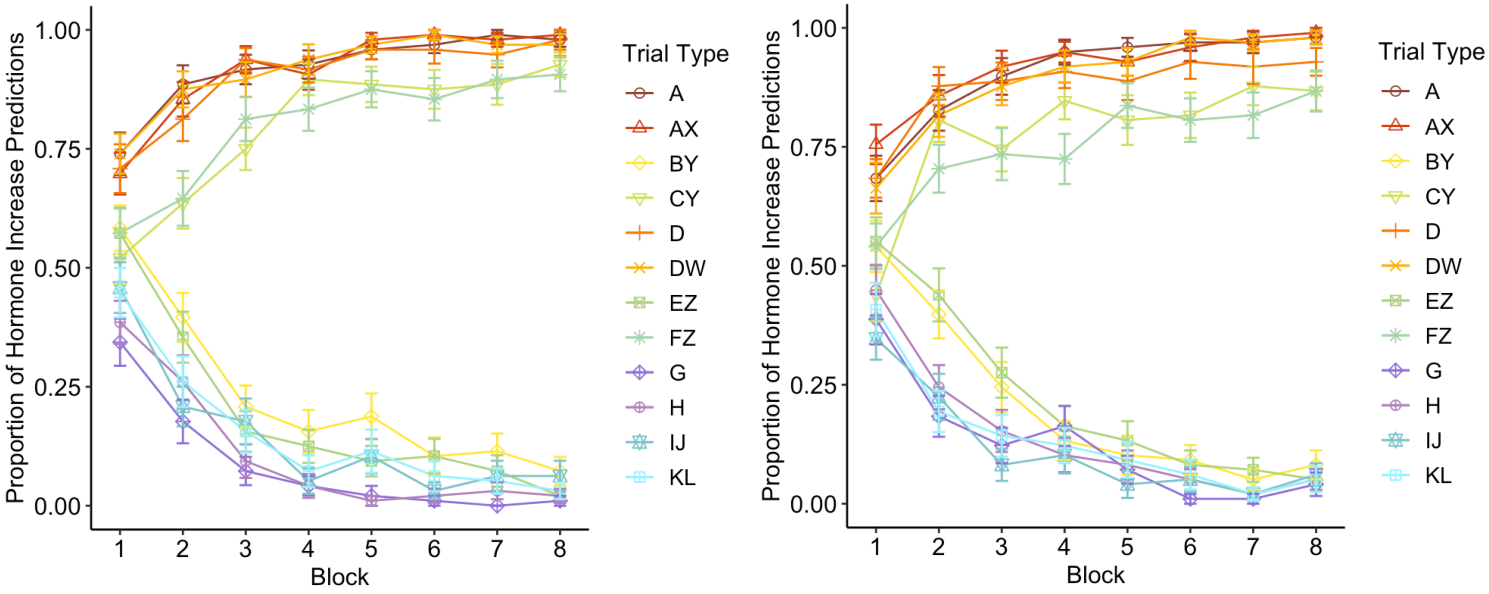
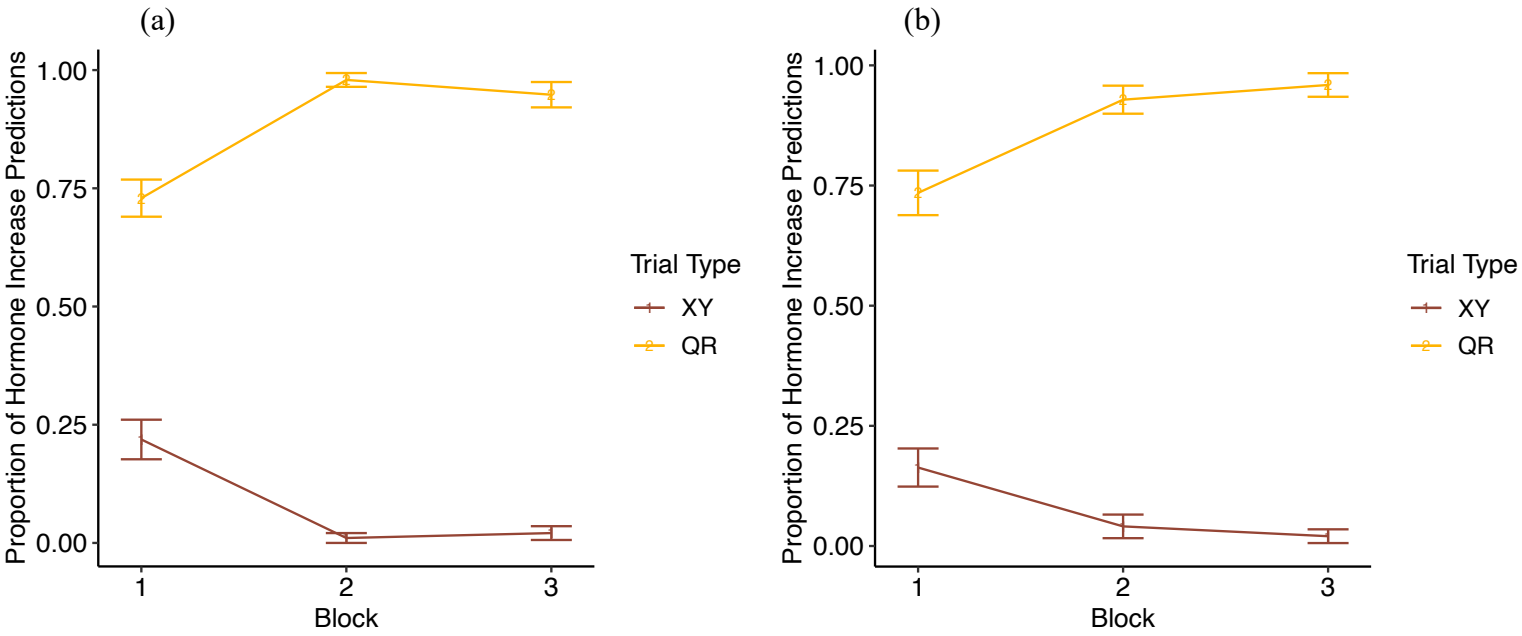


Figure S.39

Proportion of hormone increase predictions across blocks of Stage 2 compound training in Experiment 4.3 for (a) the additive group and (b) the non-additive group. Error bars indicate standard error of mean (SEM).



Experiment 4.4

Figure S.40

Proportion of hormone increase predictions across blocks of pretraining in Experiment 4.4 for (a) the preventative group and (b) the non-preventative group. Error bars indicate standard error of mean (SEM).

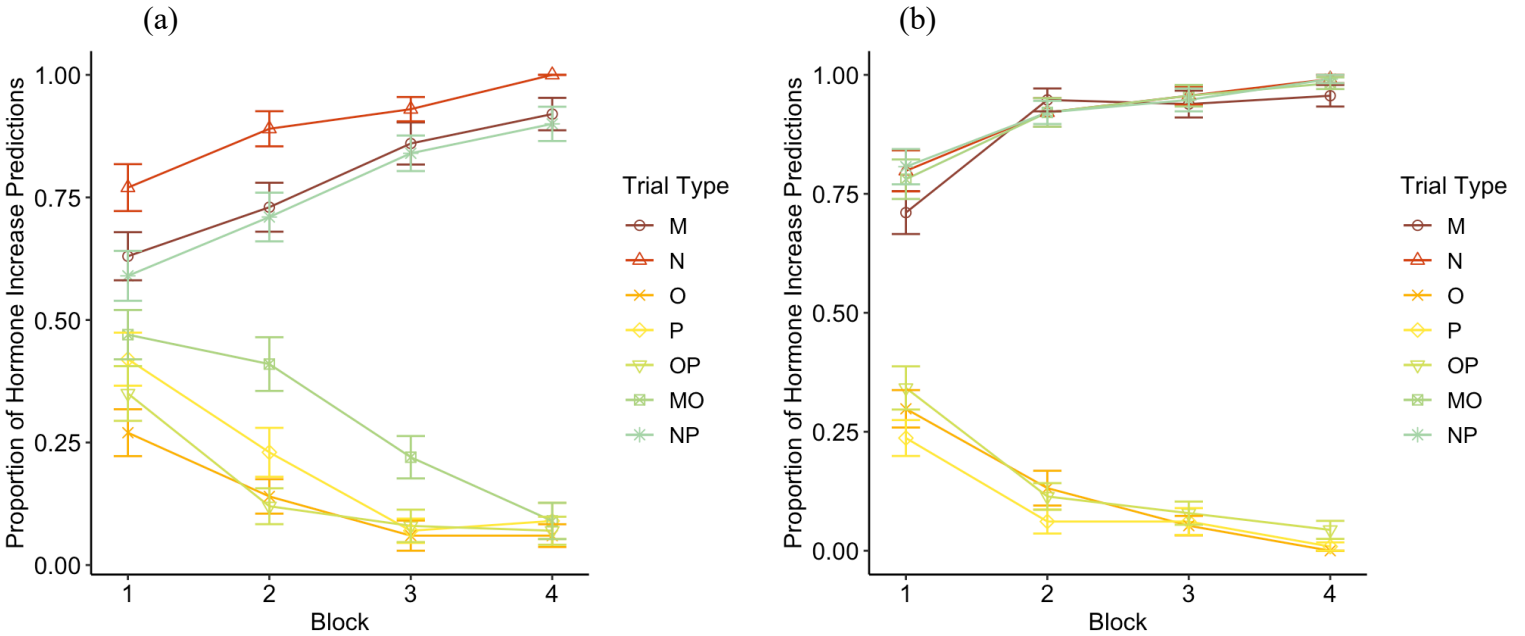


Figure S.41

Proportion of hormone increase predictions across blocks of Stage 1 training in Experiment 4.4 for (a) the preventative group and (b) the non-preventative group. Error bars indicate standard error of mean (SEM).

(a) (b)

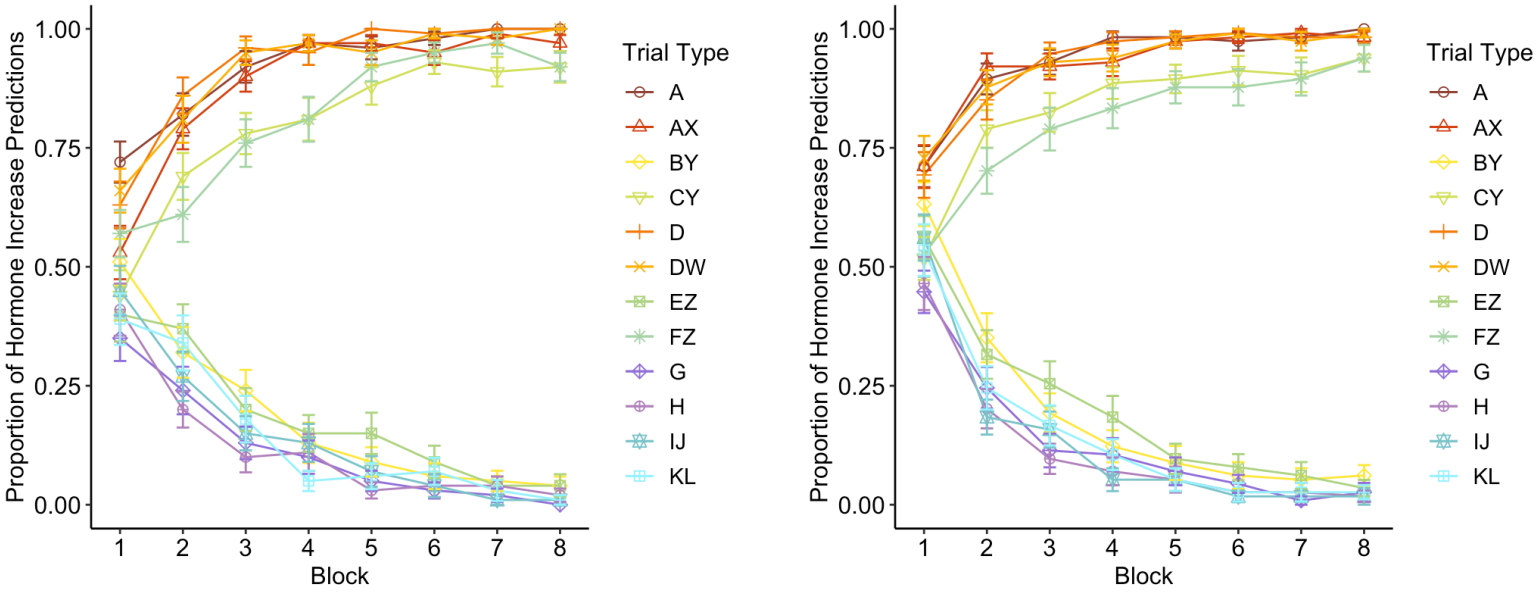
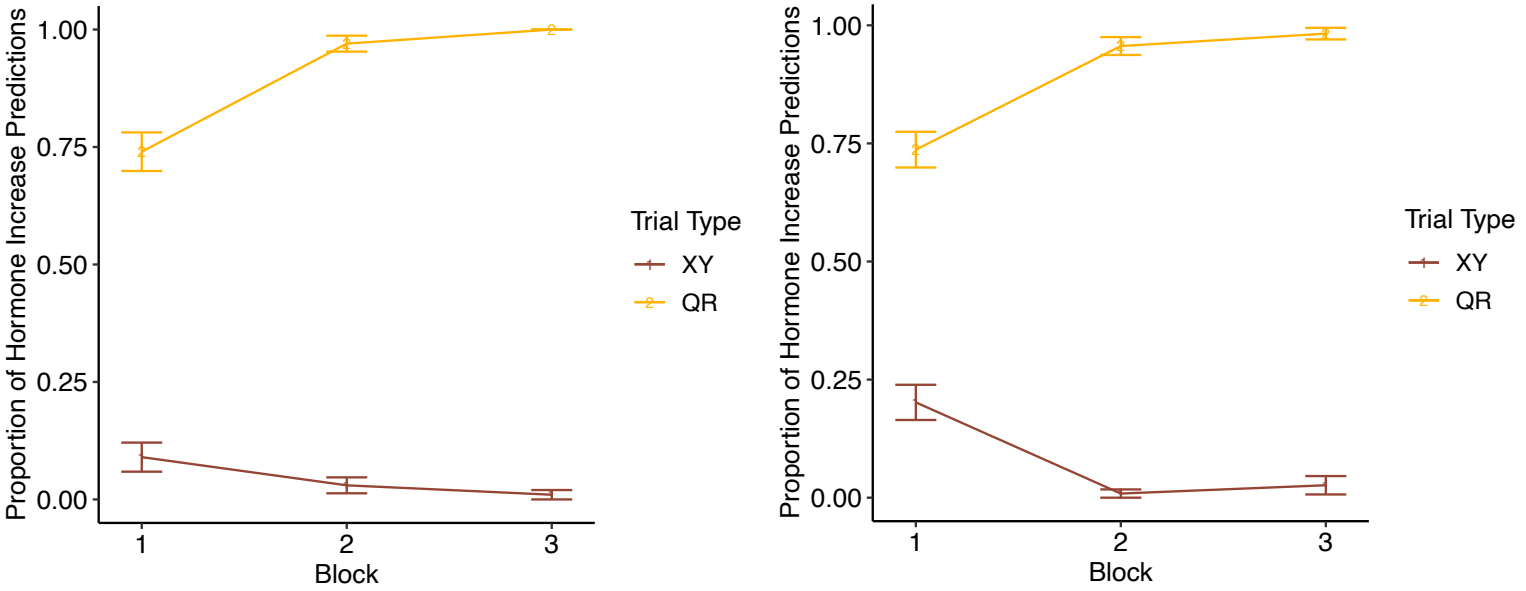


Figure S.42

Proportion of hormone increase predictions across blocks of stage 2 compound training in Experiment 4.4 for (a) the preventative group and (b) the non-preventative group.

(a) (b)



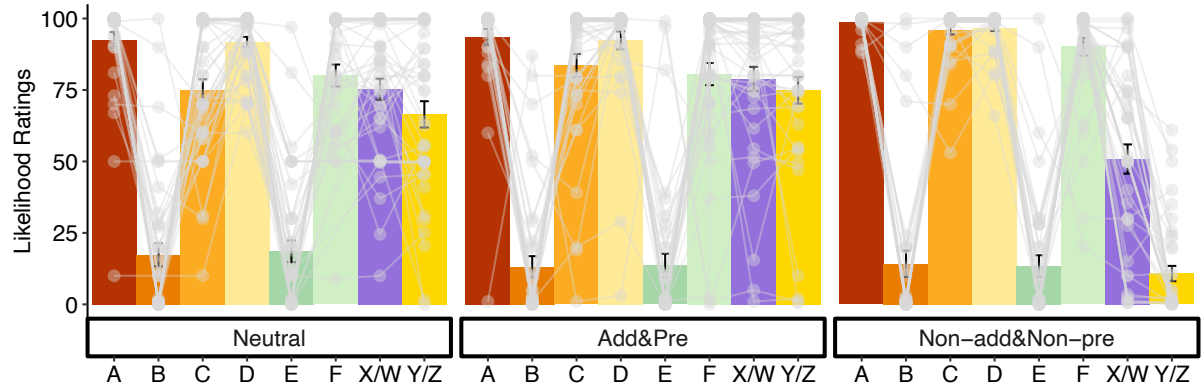
Ratings Test

Experiment 4.1

Figure S.43

(a) Mean likelihood ratings and (b) Mean confidence ratings for the neutral pretraining group, the additive and preventative group, and the non-additive and non-preventative group on the Stage 1 ratings test of Experiment 4.1. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

(a)



(b)

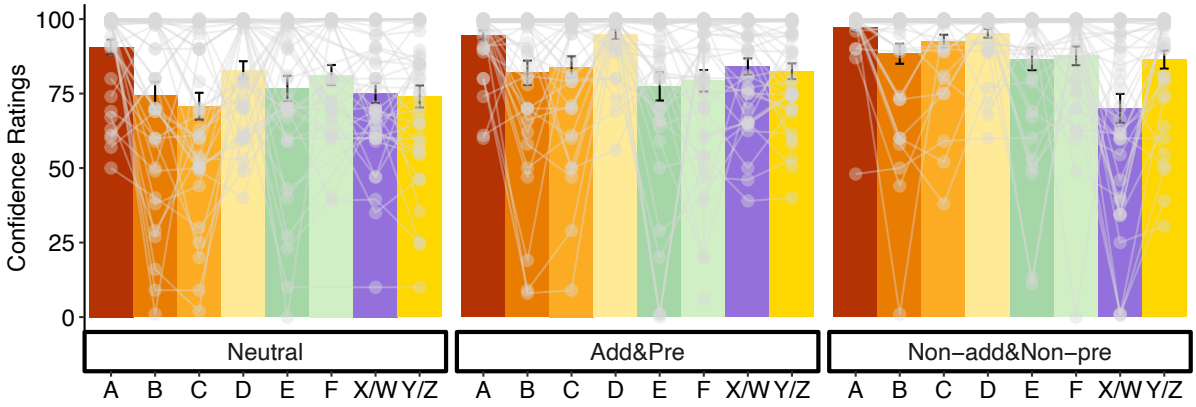
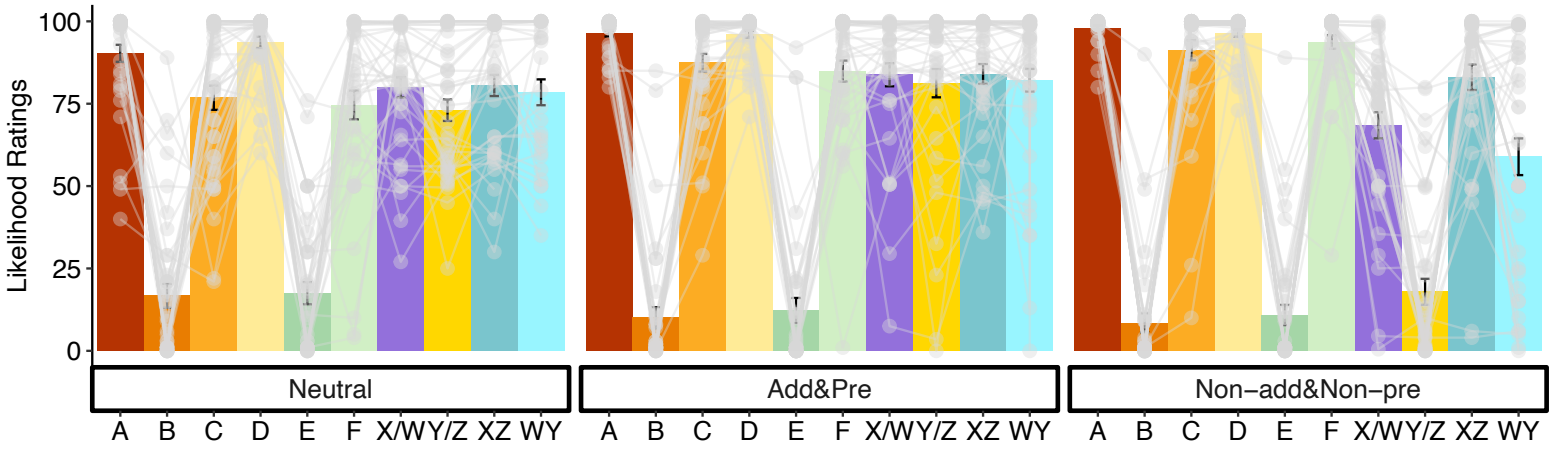


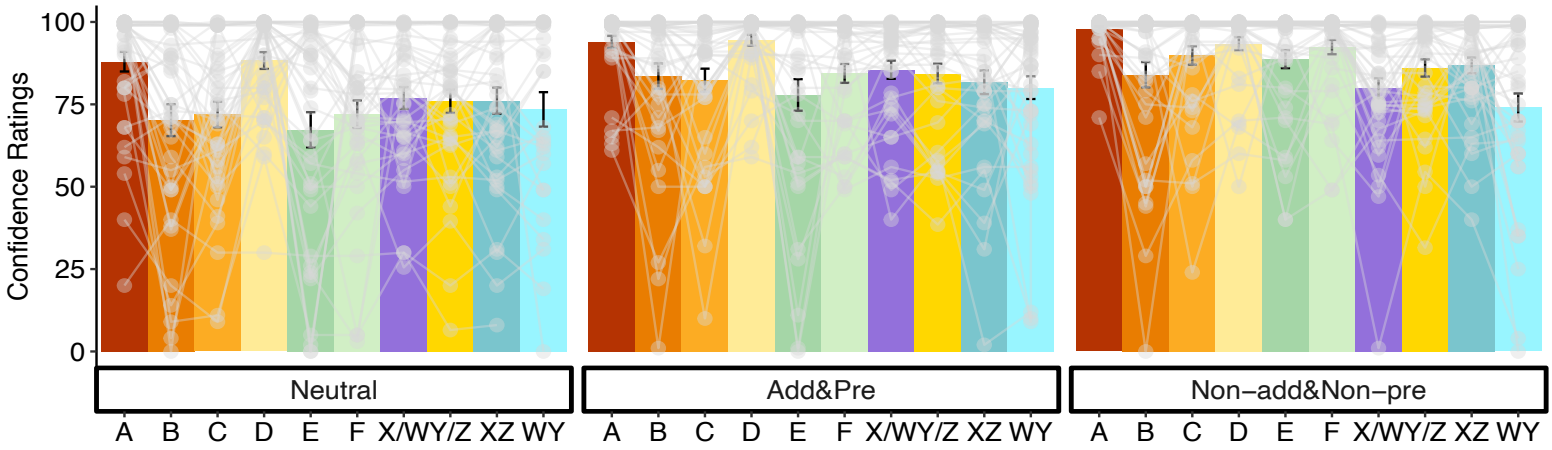
Figure S.44

(a) Mean likelihood ratings and (b) Mean confidence ratings for the neutral pretraining group, the additive and preventative group, and the non-additive and non-preventative group on the Stage 2 ratings test of Experiment 4.1. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

(a)



(b)

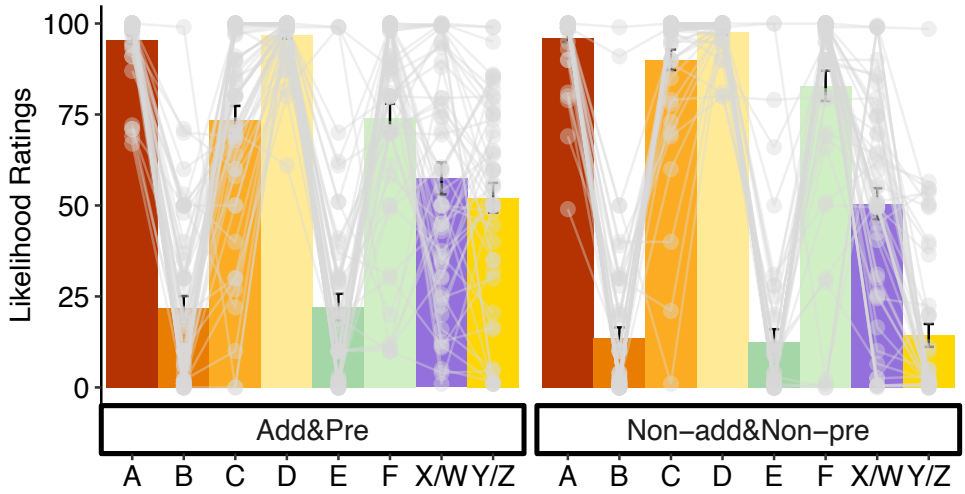


Experiment 4.2

Figure S.45

(a) Mean likelihood ratings and (b) Mean confidence ratings for the additive and preventative group, and the non-additive and non-preventative group on the Stage 1 ratings test of Experiment 4.2. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

(a)



(b)

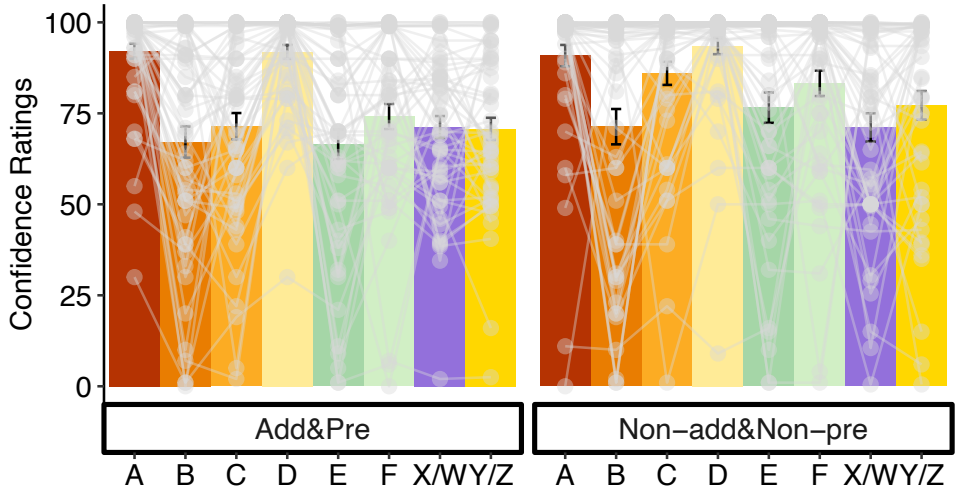
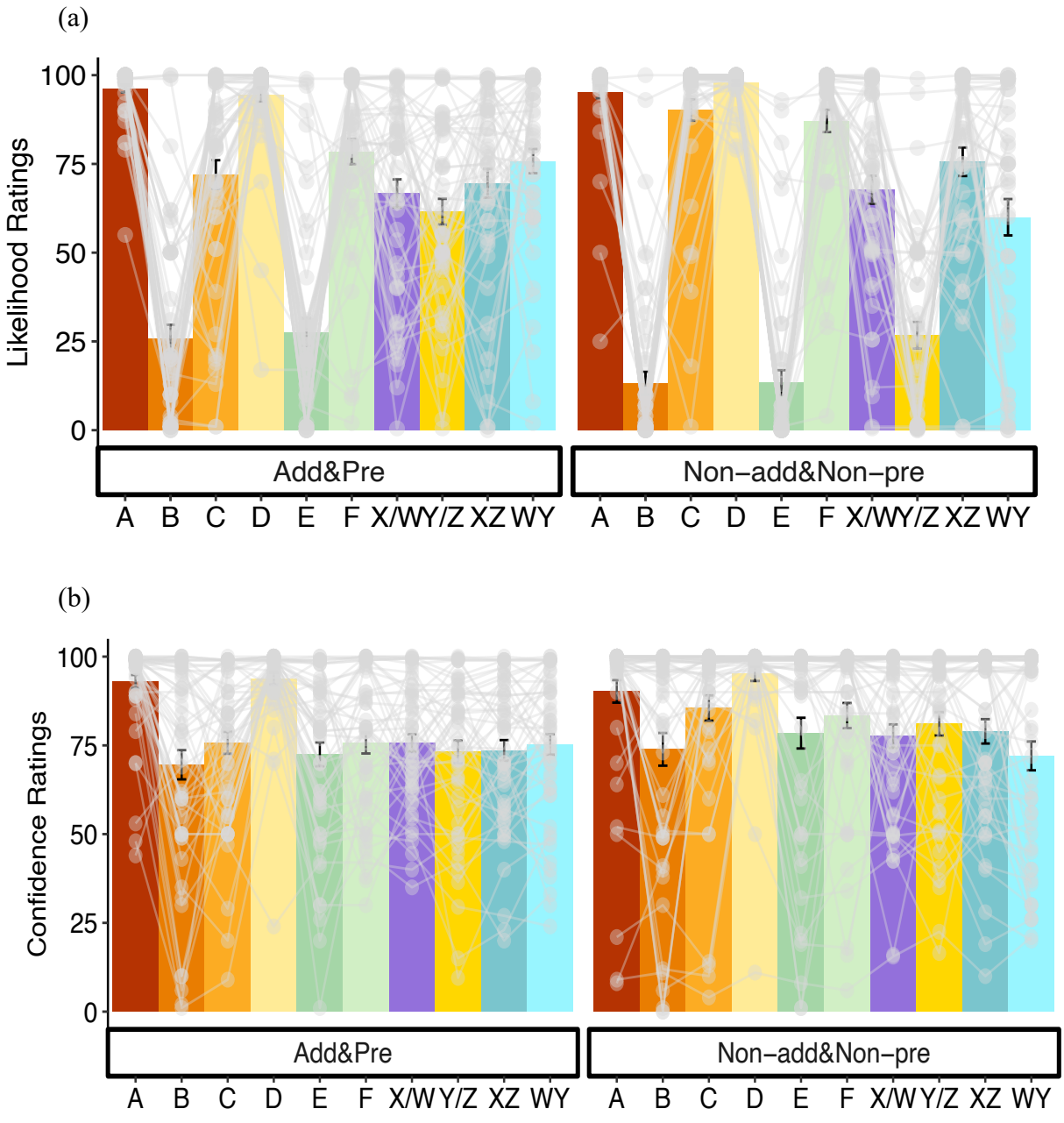


Figure S.46

(a) Mean likelihood ratings and (b) Mean confidence ratings for the additive group and the non-additive group on the Stage 2 ratings test of Experiment 4.2. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

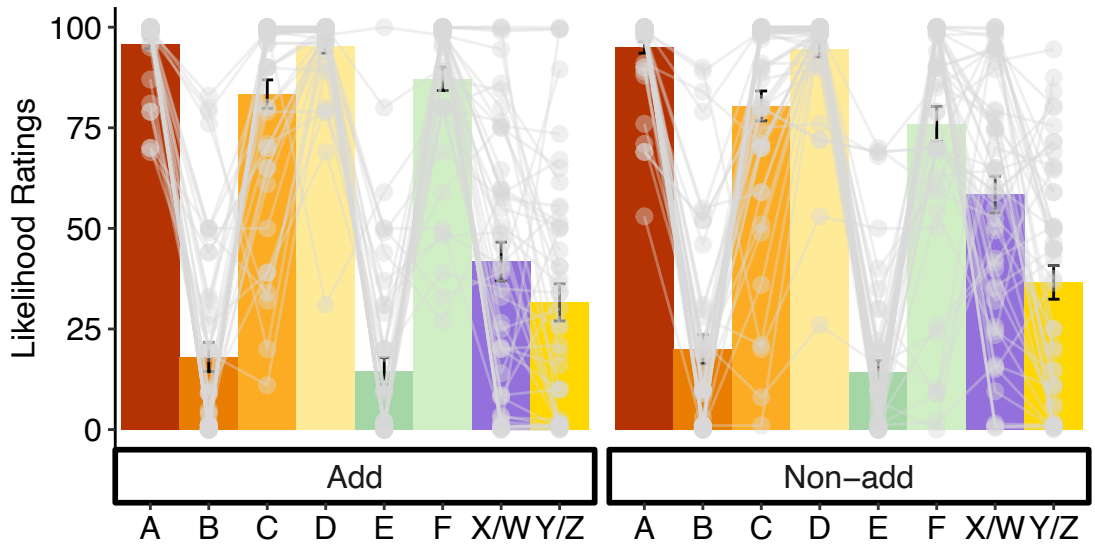


Experiment 4.3

Figure S.47

(a) Mean likelihood ratings and (b) Mean confidence ratings for the additive group and the non-additive group on the Stage 1 ratings test of Experiment 4.3. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

(a)



(b)

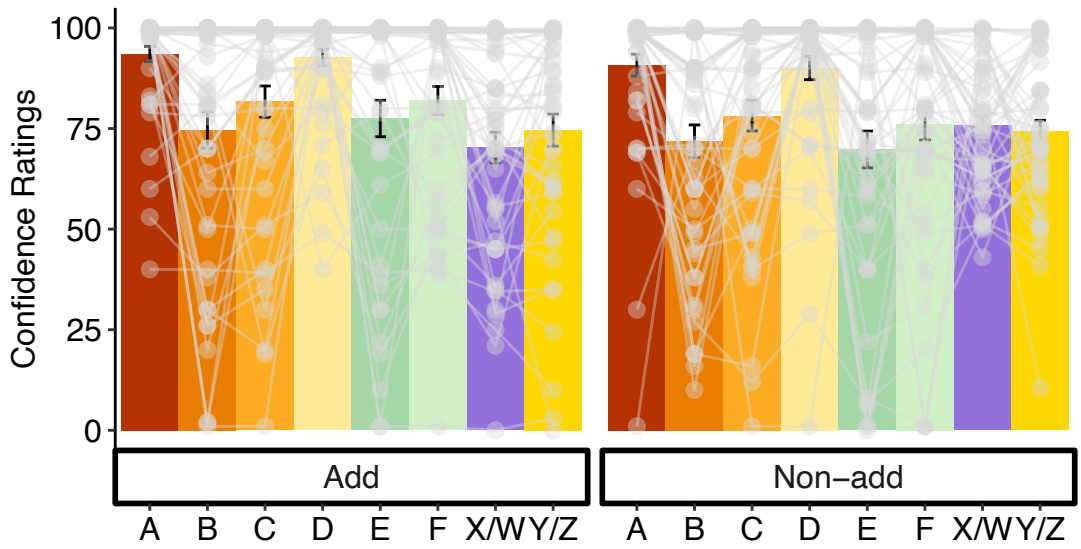
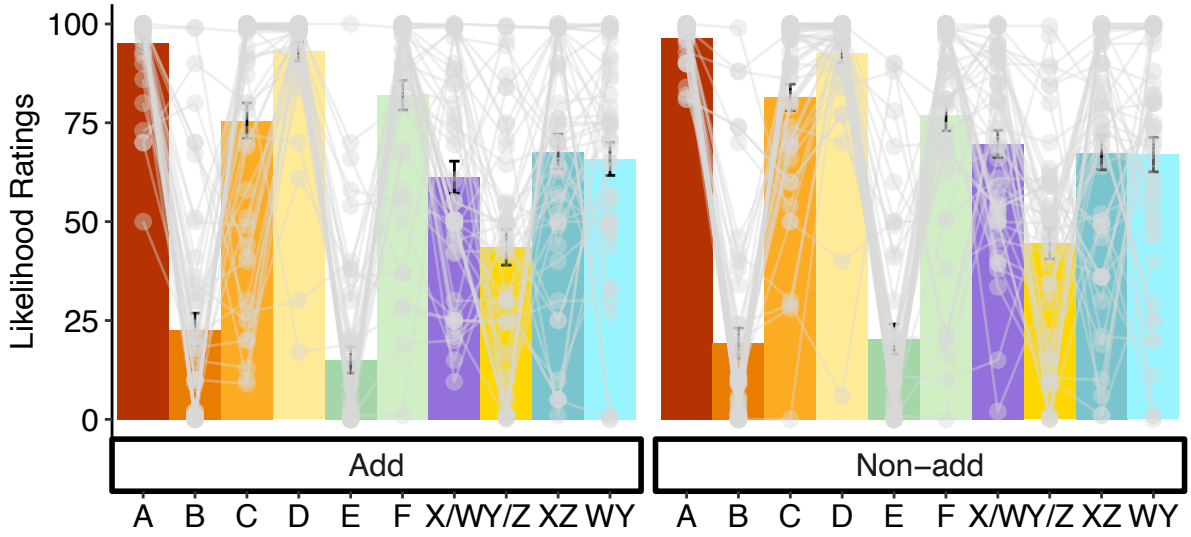


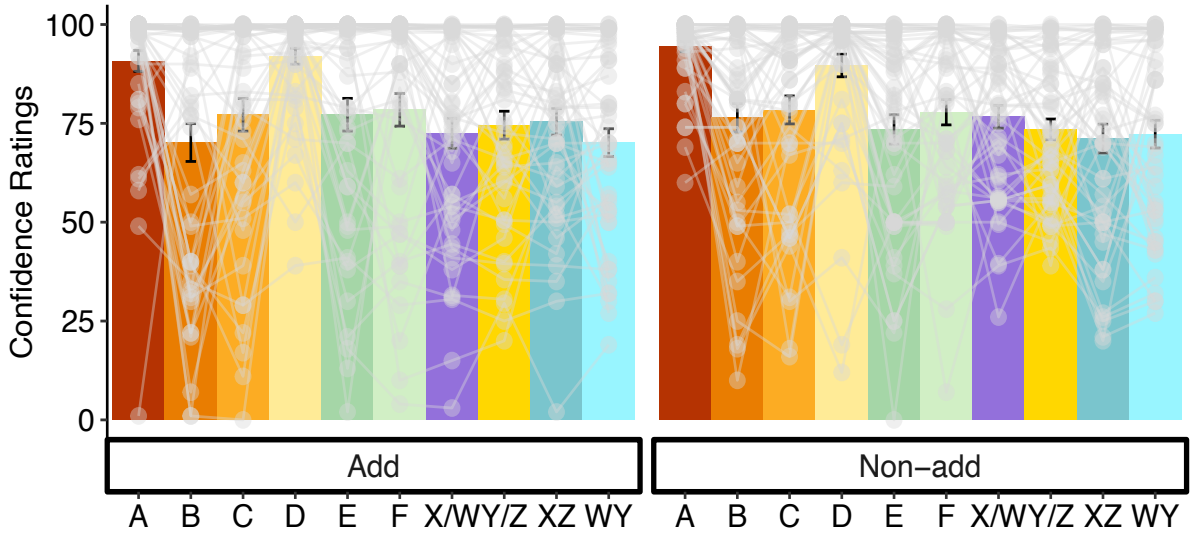
Figure S.48

(a) Mean likelihood ratings and (b) Mean confidence ratings for the additive group and the non-additive group on the Stage 2 ratings test of Experiment 4.3. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

(a)



(b)

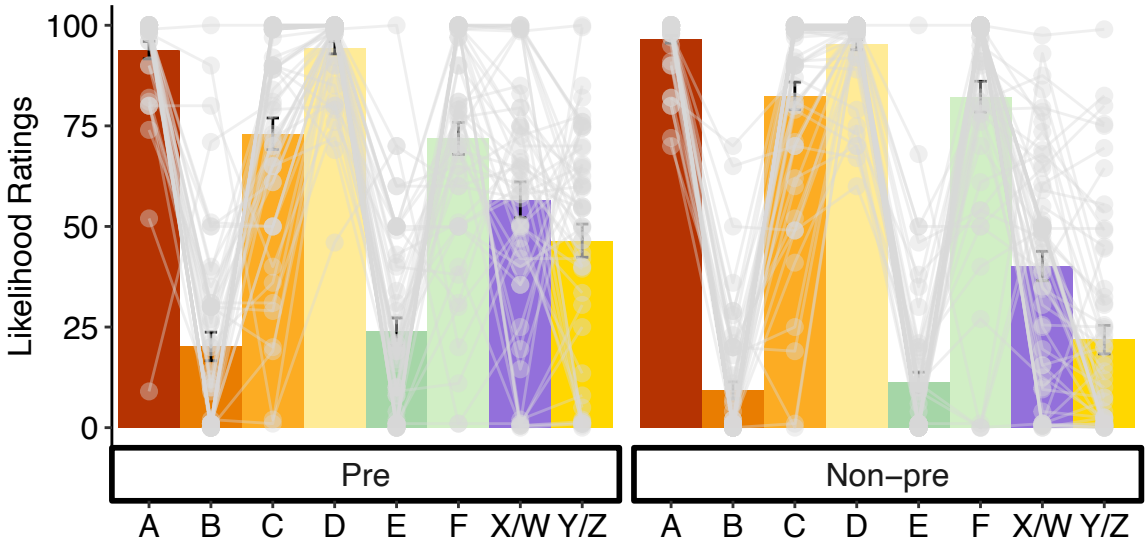


Experiment 4.4

Figure S.49

(a) Mean likelihood ratings and (b) Mean confidence ratings for the preventative group and the non-preventative group on the Stage 1 ratings test of Experiment 4.4. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

(a)



(b)

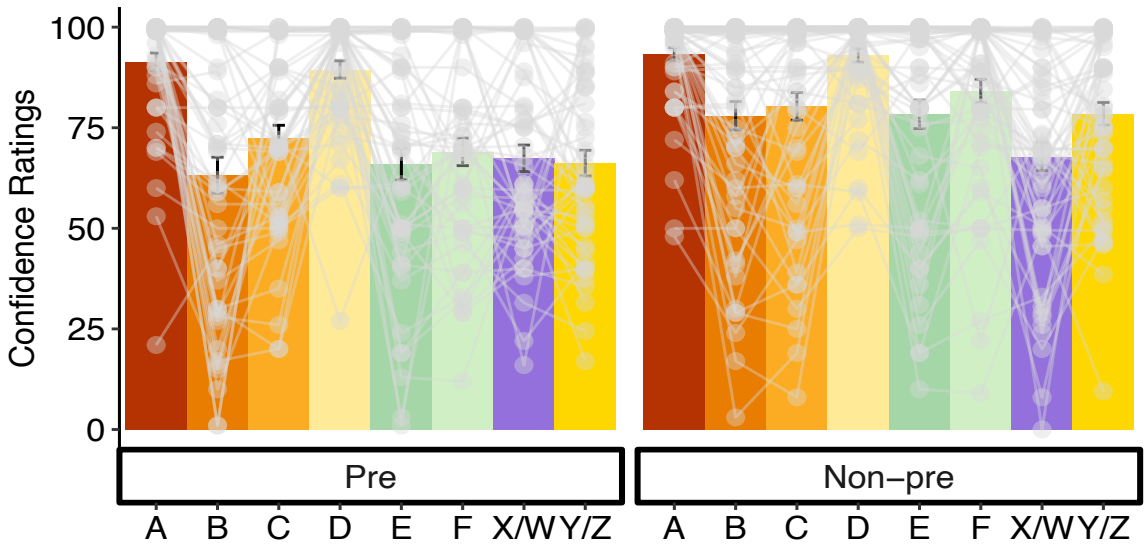
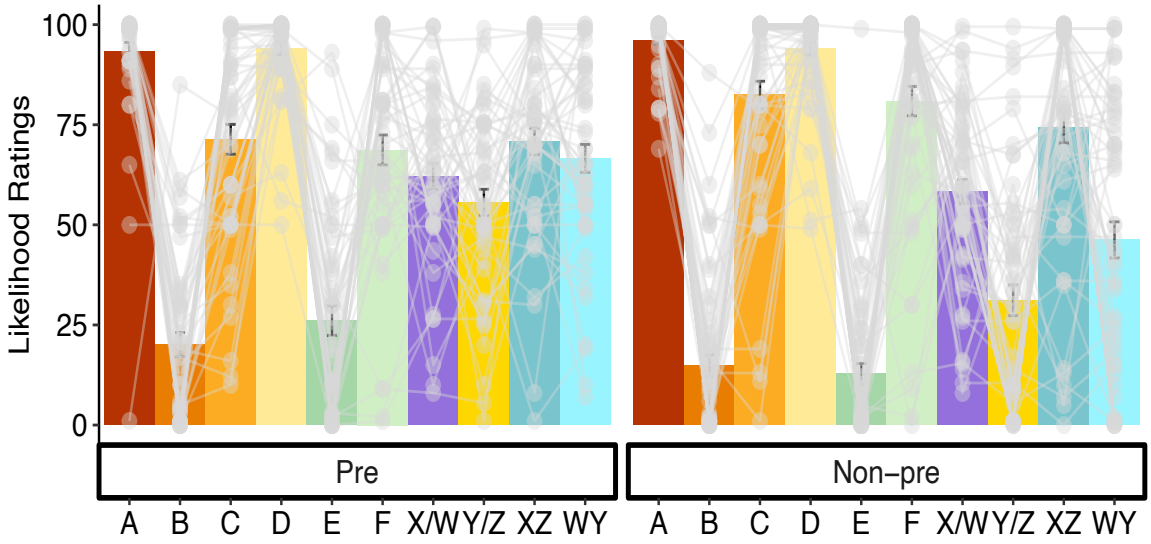


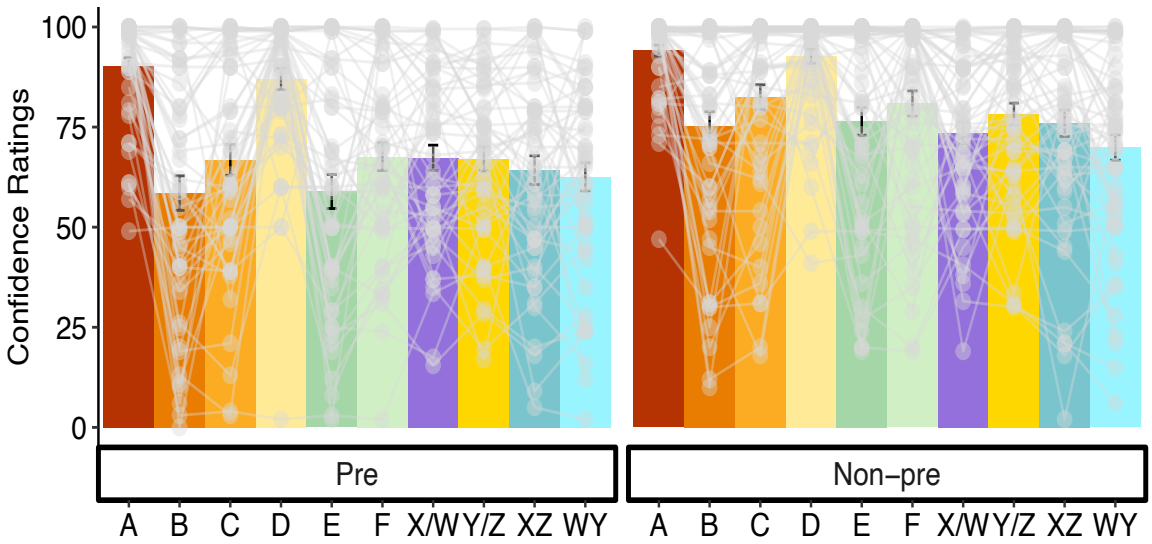
Figure S.50

(a) Mean likelihood ratings and (b) Mean confidence ratings on the Stage 2 ratings test for the neutral pretraining group, the additive and preventative group, and the non-additive and non-preventative group in Experiment 4.4. Error bars indicate standard error of mean (SEM). Connected points represent data from the same participant.

(a)



(b)

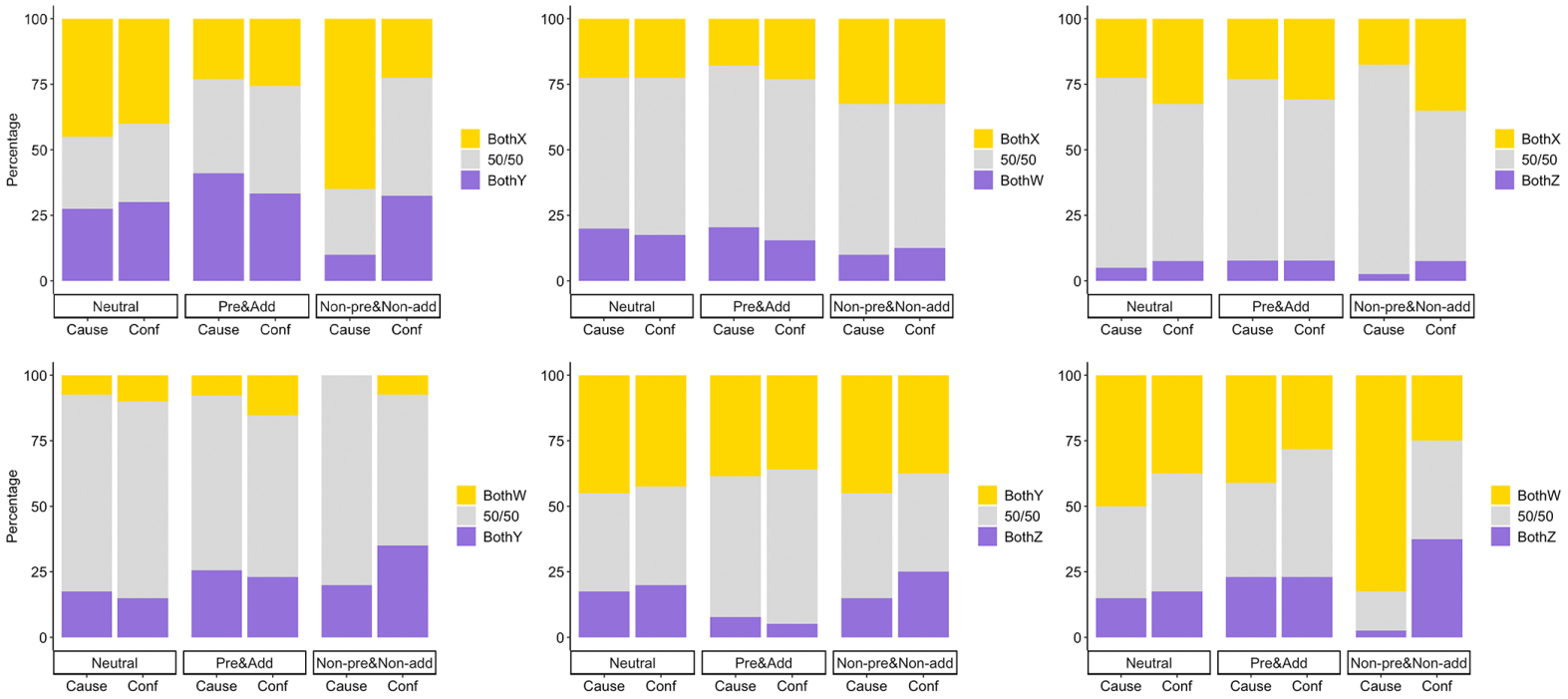


Forced Choice Test

Experiment 4.1

Figure S.51

Mean percentage of choosing the first cue from the XY pair, the XW pair, the XZ pair, the WY pair, the YZ pair, and the WZ pair on the forced choice test in Experiment 4.1. Higher percentage indicates higher likelihood that a given cue is chosen as a cause.



Forced Choice Test. The pattern of data produced by dichotomous dependent variables (i.e. selecting one cue from each cue pair) violates the gaussian distribution underlying common parametric tests. A parameter-free one-sided sign test was thus chosen for the complementary forced choice test. Choice proportions are summarised in Figure 4. The comparisons of interest were between the two blocked cues (X vs. W), the two uncorrelated cues (Y vs. Z), and within four pairs of blocked and uncorrelated cues (X vs. Y; W vs. Z[1]). For the two blocked cues, results indicated that only non-additive and non-preventative participants chose X significantly more frequently than W, p=.025, and did so with higher confidence for X, p=.048. For the two uncorrelated cues, participants across all

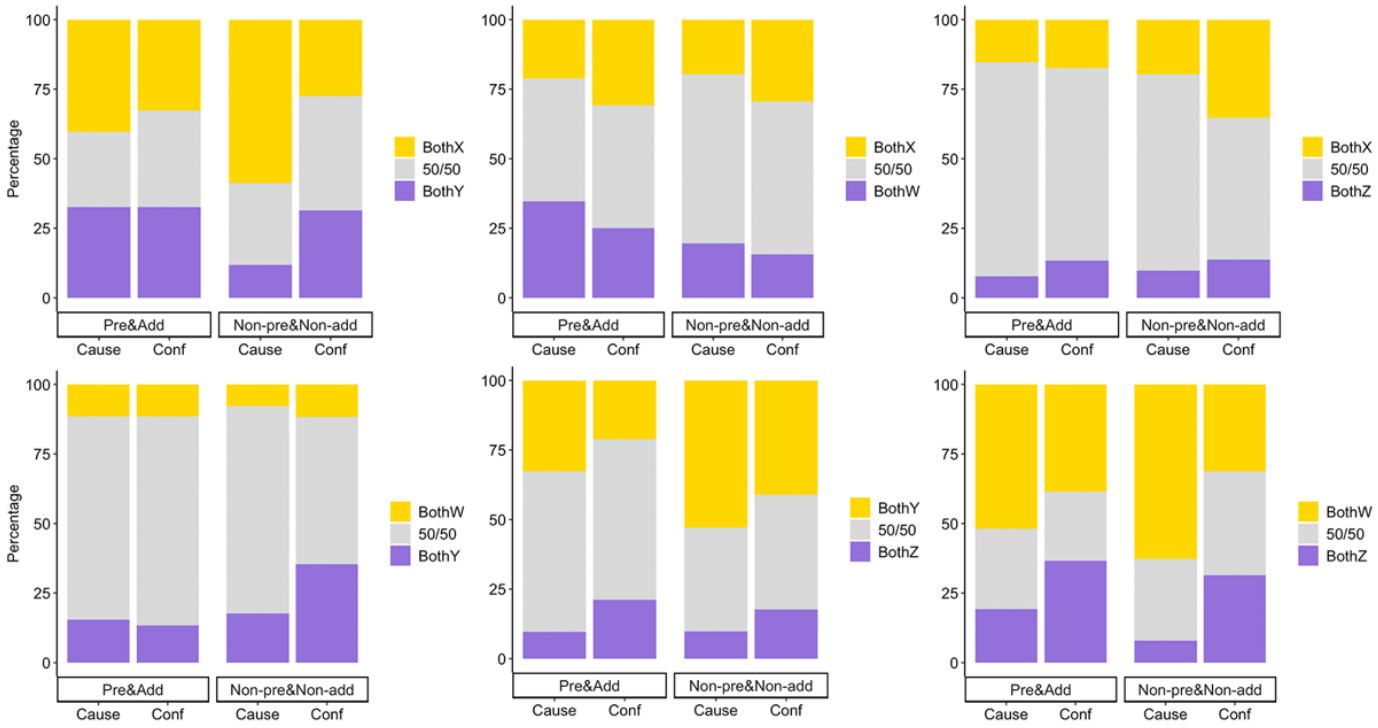
three groups were inclined to choose Y as the more likely cause than Z, where p=.004 for the additive and preventative group, p=.011 for the non-additive and non-preventative group, and p=.021 for the neutral group. In particular, the additive and preventative group made this judgment with higher confidence for Y over Z, p=.002, while the non-additive and non-preventative group did not, p=.212. Although the neutral group was slightly more confident about Y than Z, this was not statistically significant, p=.053. Comparison between redundant cues revealed that non-additive and non-preventative participants were more likely to choose X over Y, p<.001, and W over Z, p<.001; and neutral participants were more likely to choose W over Z, p=.005, as being the more probable causes.

Experiments 4.2–4.4

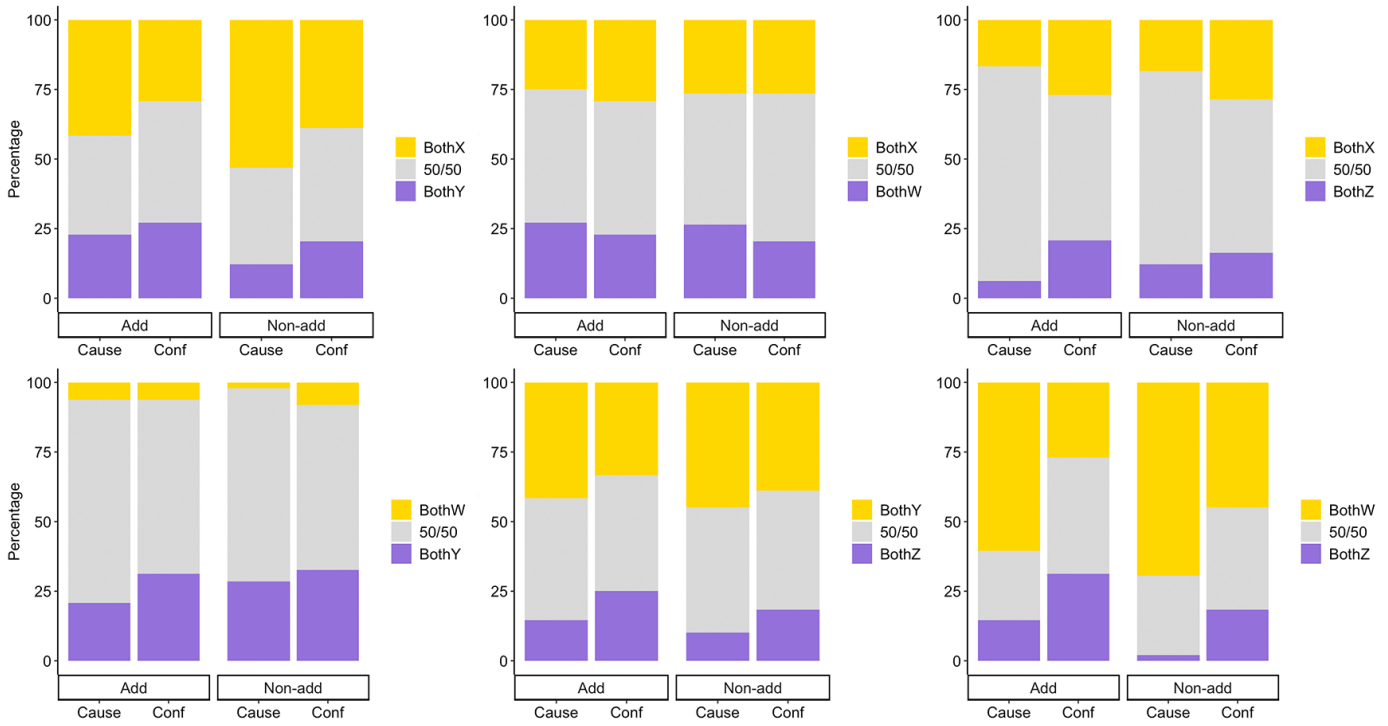
Figure S.52

Mean percentage of choosing the first cue from the XY pair, the XW pair, the XZ pair, the WY pair, the YZ pair, and the WZ pair on the forced choice test in Experiment 4.2-4.4. Higher percentage indicates higher likelihood that a given cue is chosen as a cause.

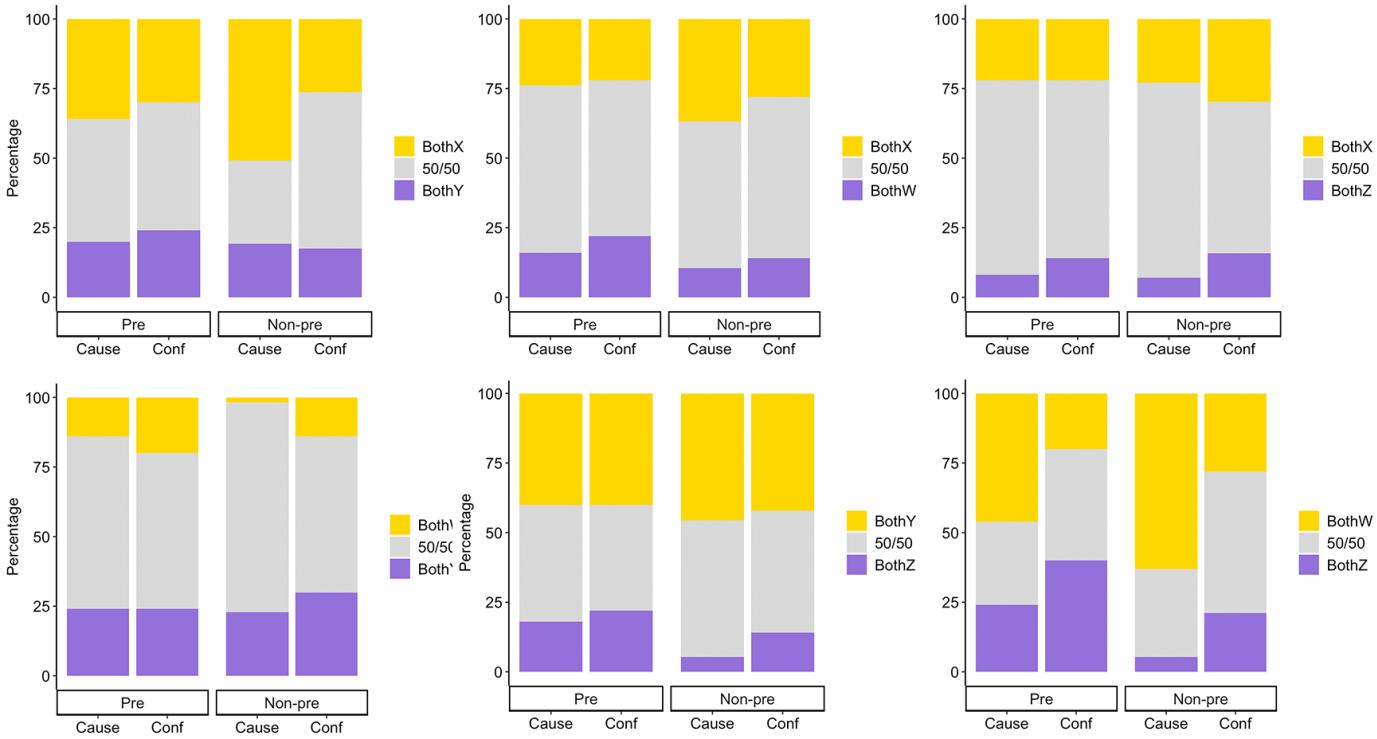
(a) Experiment 4.2



(b) Experiment 4.3



(c) Experiment 4.4



Forced Choice Test. Choice proportions on the forced choice test are shown in Figure X. One-sided sign tests showed that the additive and preventative group had a higher probability of choosing Z as the less likely cause than Y, p=.008, or W, p=.004. The non-additive and non-preventative group on the other hand chose Y as the more likely cause with higher probability over Z, p<.001, chose X with higher probability over Y, p<.001, and chose W with higher probability over Z, p<.001. They also chose Y over Z, p=.021, X over Z, p=.022, and Y over W, p=.011, as the cues for which they had more confidence.

Figure X illustrates choice percentages on the forced choice test. Sign test results for the additive group indicated significantly higher choice proportions in terms of causal likelihood for Y over Z, p=.009, for Y over W, p=.046, and for W over Z, p<.001. In particular, Y was chosen as the more confident cue than Z with significantly higher probability, p=.004. Results for the non-additive group revealed significantly higher choice proportions in terms of causal likelihood for Y over Z, p<.001, for X over Y, p<.001, for Y over W, p<.001, and for W over Z, p<.001. Non-additive participants also chose Y over Z, p=.044, W over Z, p=.015, and Y over W, p=.006, as the cue they felt more confident judging with significantly higher probability.

Proportion of choices for each pair of comparison on the cue choice test is shown in Figure X. The sign test results revealed that preventative participants chose Y significantly more frequently over Z as the more probable cause, p=.031. Preventative participants also chose W more frequently over Z as the more probable cause, p=.045, and as the more confident cue, p=.049. Non-preventative participants had a significantly higher probability of choosing X over W, p=.003, Y over Z, p<.001, X over Y, p=.003, X over Z, p=.025, Y over W, p<.001, and W over Z, p<.001, as the more likely causes. Among these choice biases, preventative participants also chose Y more often than Z as the cue they felt more confident judging, p=.004.

Supplementary Materials: Chapter 5

Summary Descriptions

The Food Allergist Scenario

Imagine that you are an allergist trying to determine the cause of an allergic reaction shortly after your patient eats a meal. You arrange that the patient to eat particular foods over a series of consecutive days, and then report to you whether an allergic reaction occurred. Sometimes a single food is eaten, and sometimes a combination of two foods is eaten together. The following summary describes what foods the patient has consumed each day and whether there was an allergic reaction.

The Redundancy Effect.

On 10 separate days, the patient ate Food A and each time experienced an ALLERGIC REACTION.

On 10 separate days, the patient ate Food A and Food B, and each time experienced an ALLERGIC REACTION.

On 10 separate days, the patient ate Food C and Food D, and each time experienced an ALLERGIC REACTION.

On 10 separate days, the patient ate Food D and Food E, and each time experienced NO REACTION.

On 10 separate days, the patient ate Food F and each time experienced NO REACTION.

On 10 separate days, the patient ate Food F and Food G, and each time experienced NO REACTION.

The Relative Validity Effect.

On 10 separate days, the patient ate Food A and Food B, and each time experienced an ALLERGIC REACTION.

On 10 separate days, the patient ate Food B and Food C, and each time experienced NO REACTION.

On 10 separate days, the patient ate Food D and Food E, and experienced an ALLERGIC REACTION on HALF of the days.

On 10 separate days, the patient ate Food D and Food F, and experienced an ALLERGIC REACTION on HALF of the days.

The Hormone Change Scenario

Imagine that you are a medical researcher trying to determine the effect of various medicines on a patient's hormone levels. You arrange that the patient take particular medicines over a series of consecutive days, and then record whether a hormone change occurred. Sometimes a single medicine is administered, and sometimes a combination of two medicines is administered together. Each medicine could cause an increase in hormone levels, have no impact on hormone levels, or could prevent a hormone increase that would otherwise occur. The following summary describes what medicines the patient has taken each day and whether there was a hormone change.

The Redundancy Effect.

On 10 separate days, the patient took Medicine A and each time experienced an INCREASE in hormone levels.

On 10 separate days, the patient took Medicine A and Medicine B, and each time experienced an INCREASE in hormone levels.

On 10 separate days, the patient took Medicine C and Medicine D, and each time experienced an INCREASE in hormone levels.

On 10 separate days, the patient took Medicine D and Medicine E, and each time experienced NO CHANGE in hormone levels.

On 10 separate days, the patient took Medicine F and each time experienced NO CHANGE in hormone levels.

On 10 separate days, the patient took Medicine F and Medicine G, and each time experienced NO CHANGE in hormone levels.

The Relative Validity Effect.

On 10 separate days, the patient took Medicine A and Medicine B, and each time experienced an INCREASE in hormone levels.

On 10 separate days, the patient took Medicine B and Medicine C, and each time experienced NO CHANGE in hormone levels.

On 10 separate days, the patient took Medicine D and Medicine E, and experienced an INCREASE in hormone levels on HALF of the days.

On 10 separate days, the patient took Medicine D and Medicine F, and experienced an INCREASE in hormone levels on HALF of the days.

Rating Test

Figure S.53

(a) Mean likelihood ratings and (b) Mean confidence ratings for the redundancy effect tested in the food allergist task and the hormone change task with counterbalanced orders in Experiment 5.1. The middle two pairs of bars illustrate the first set of ratings made by participants (i.e. medicine ratings for the hormone change task first group and food ratings for the food allergist task first group). Dots indicate scores from individual participants and error bars indicate standard error of mean (SEM).

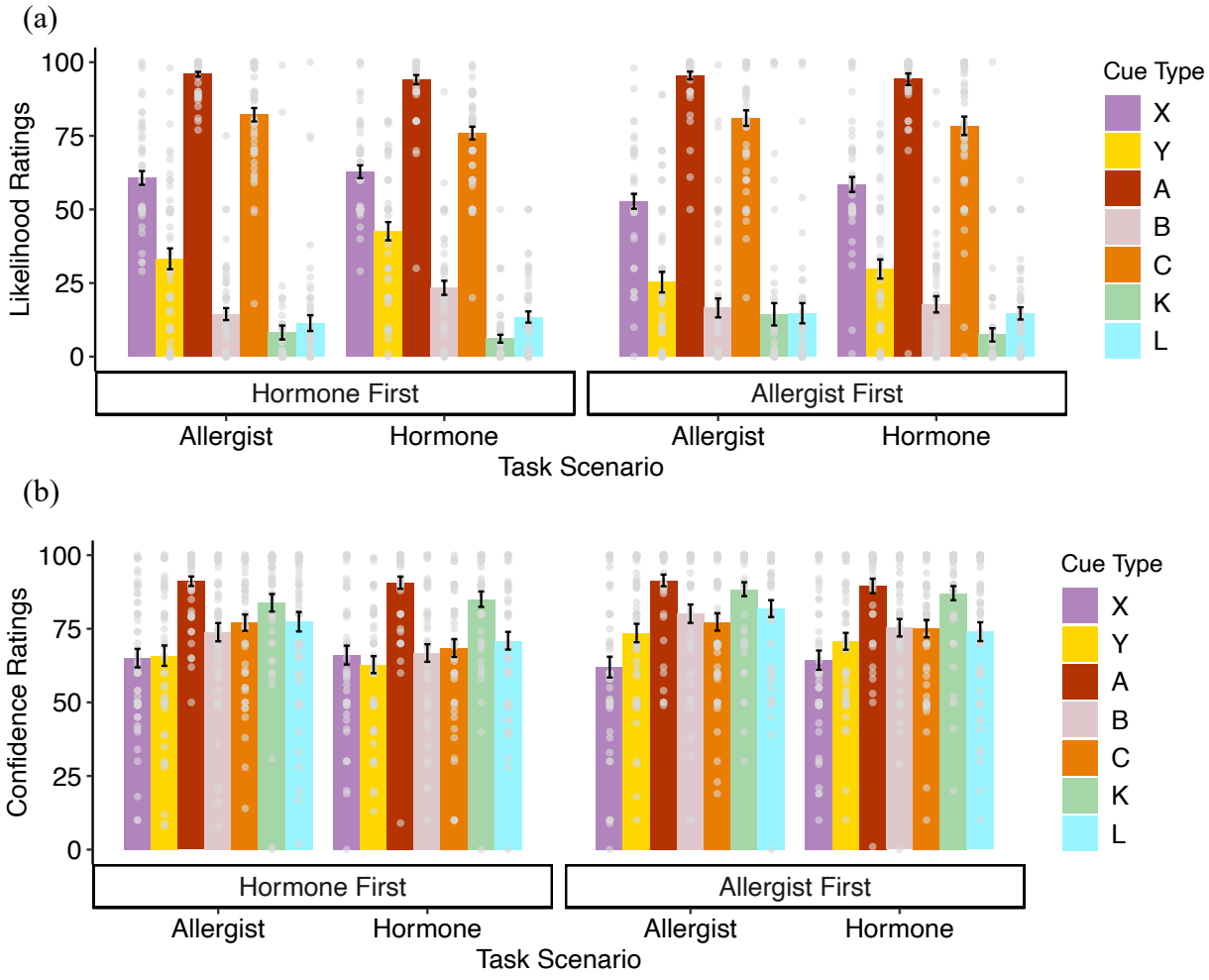
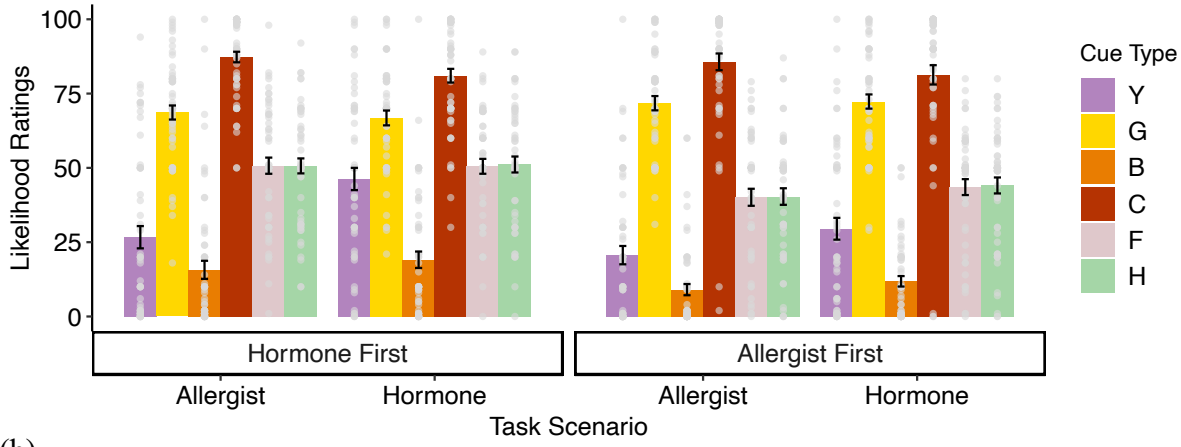


Figure S.54

(a) Mean likelihood ratings and (b) Mean confidence ratings for the relative validity effect tested in the food allergist task and the hormone change task with counterbalanced orders in Experiment 5.2. Dots indicate scores from individual participants and error bars indicate standard error of mean (SEM).

(a)



(b)